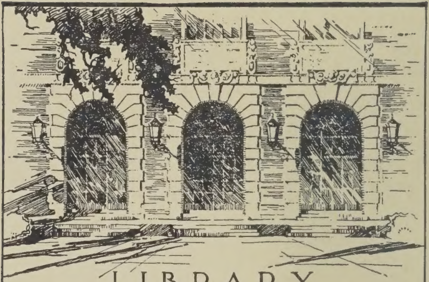


LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

[redacted]

Allerton Conference on Circuit and
System Theory 1964

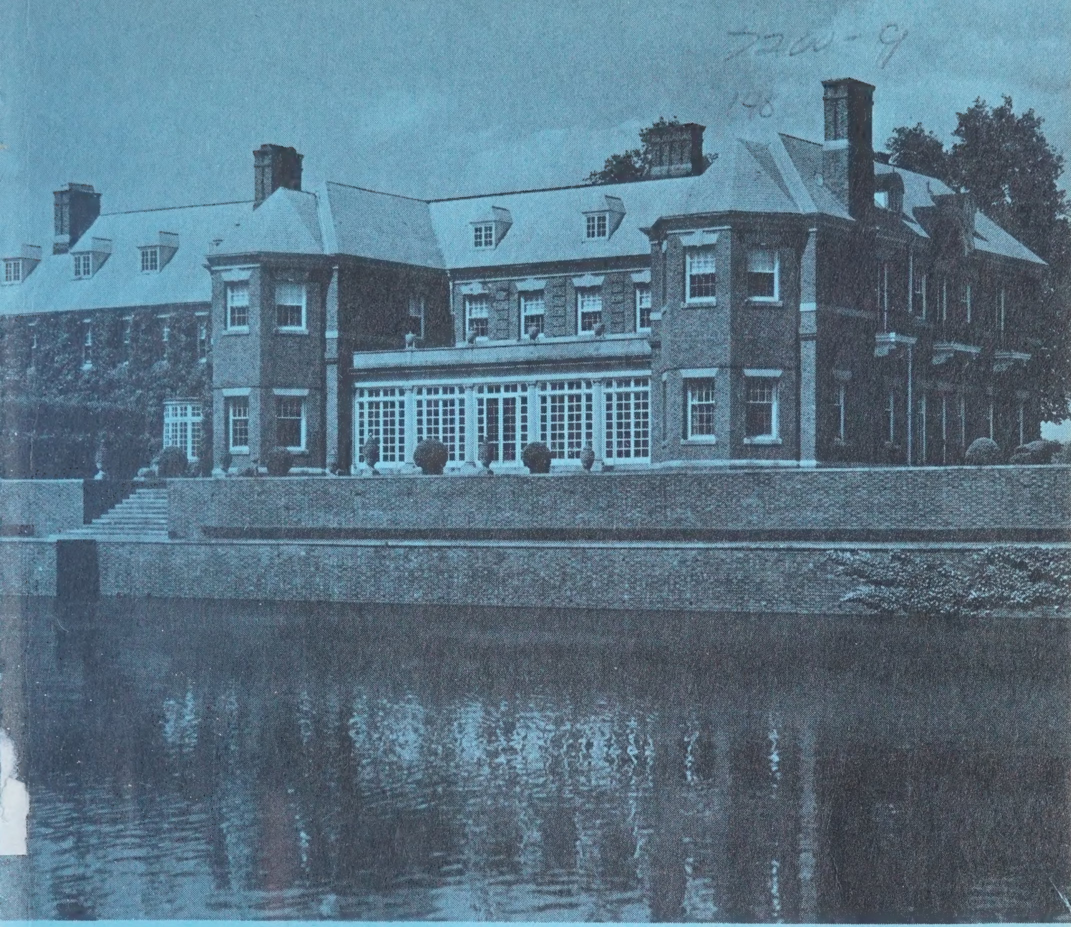
Proceedings

✓

Digitized by the Internet Archive
in 2026 with funding from
University of Illinois Urbana-Champaign

https://archive.org/details/proceedings2196alle_0

7200-9
148



Proceedings

SECOND ANNUAL ALLERTON CONFERENC ON CIRCUIT AND SYSTEM THEORY

DIGITAL COMPUTER
LABORATORY
LIBRARY

September 28-30, 1964
University of Illinois
Allerton House, Monticello, Illinois

Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois, Urbana, Illinois

OCT 25 1968

PROCEEDINGS
SECOND ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

DIGITAL COMPUTER
LABORATORY
LIBRARY

edited by

W. R. PERKINS
Conference Chairman

PATRICIA M. HINOJOSA
Production Editor

Conference held

September 28, 29, 30, 1964
Allerton House
Monticello, Illinois

Sponsored by the

DEPARTMENT OF ELECTRICAL ENGINEERING
and the
COORDINATED SCIENCE LABORATORY
of the
UNIVERSITY OF ILLINOIS
Urbana, Illinois

FOREWORD

The second annual Allerton Conference on Circuit and System Theory was held at Allerton House from September 28 to 30, 1964. A wide range of topics concerned with the related areas of circuits, networks, and systems was discussed. The papers given at this conference, together with Discussion Remarks submitted by some of the conference participants, are presented in this Proceedings.

An important feature of the Allerton Conference is the opportunity for vigorous interchange of ideas. This is due in part to the relatively small attendance and to the isolation of Allerton Park. We are pleased to note that new results and interests have developed from discussions at Allerton.

We acknowledge encouragement given by Professor Daniel Alpert, Director of the Coordinated Science Laboratory and Professor E. C. Jordan, Head of the Department of Electrical Engineering at the University of Illinois. Local arrangements and registration were handled by Mr. H. B. Lawler and Miss Rosa Cornell of the Department of Electrical Engineering. Publication of these Proceedings was handled by Mr. R. A. Kingery, Director of Engineering Publications, and Mrs. Patricia M. Hinojosa, Assistant Editor of Engineering Publications, College of Engineering, University of Illinois.

William R. Perkins
M. E. Van Valkenburg

TABLE OF CONTENTS

FOREWARD iii

GENERAL SESSION I 1

Chairman: W. R. Perkins, University of Illinois

Opening Remarks, M. E. Van Valkenburg, University of Illinois

Some Results in Stability Theory, C. A. Desoer 1

Dual Dynamical Systems and Their Representation by System Functions, L. Weiss 15

Canonical Realization, R. E. Kalman 32

CONTROL AND COMMUNICATIONS SYSTEMS 46

Chairman: R. T. Chien, IBM, Yorktown Heights, New York

The Use of Lur'e Forms to Establish a Sufficient Condition for Stability of a Class of Discrete Feedback Systems with Parallel Nonlinear Elements, G. J. Murphy and S. H. Wu 46

Minimizing Vulerability in a Communication Network, S. D. Bedrosian 70

On Flows in Communication Networks, D. T. Tang . . 83

Multi-Return Differences of Feedback Systems, G. C. Reis 100

FILTER THEORY 114

Chairman: F. F. Kuo, Bell Telephone Laboratories, Murray Hill, New Jersey

Numerical Techniques of Fourier Transforms with Applications, M. L. Liou 114

Synthesis of Cascaded Lattice Crystal-Capacitor Bandpass Filters, K. Y. Cheng 135

Determination of Characteristic Functions for Crystal-Capacitor Bandpass Ladder Filters with Nearly Chebyshev Pass Band Attenuation, M. Sablatash 149

Computer Solution of the Network Realization Problem, D. A. Calahan 175

On the Optimum Least-Squares Approximation to the Ideal Lowpass, C. Pottle and J. S. Thorp 201

The Derivation of Minimum-Phase Network Functions from Time-Delay Specifications, S. Karni and N. Ahmed 220

NONLINEAR THEORY 233
Chairman: N. Wax, University of Illinois

On the Construction of Phase-Trajectories, G. C. Reis 233

Characterization of Non-Elementary Singular Points of Nonlinear Systems in the Phase Plane, W. H. Chen 250

Second Order Autonomous, Nonlinear System: A Graphical Solution, T. A. Bickart 273

On the Existence of Normal Form for Nonlinear Networks at Large with Controlled Sources, L. V. Auth, Jr. and R. -w. Liu 288

PANEL DISCUSSION

What are the Prospects for Continuing Support of Basic Engineering Research, R. H. Wilcox, B. S. Morgan, Jr., K. A. Pullen

GENERAL SESSION II 293
Chairman: M. E. Van Valkenburg, University of Illinois

Linear Degeneracies in N-Port Networks, P. P. Civalleri 293

Computers as an Aid in the Development of Transmission Circuits, C. W. Rosenthal and H. Simon 337

The Dynamic Range of Circuits with Multiport Nonlinearities, D. R. Anderson and B. J. Leon 366

SYSTEM THEORY 377
Chairman: W. R. Perkins, University of Illinois

On System Functions with the Property of Separability, L. Weiss 377

Synthesis of Linear Time-Varying Systems, G. S. Glinski and M. Arozullah 394

On the Stability of Systems Containing A Time-Varying Gain, R. W. Brockett and L. J. Forys 413

New Properties of State Equations and Associated Higher Order Differential Equations of Linear, Time-Invariant Systems, R. H. Rogers and D. P. Brown 431

TOPOLOGICAL METHODS 447
Chairman: W. Mayeda, University of Illinois

On the Differeing Abilities of RL Structures to Realize Natural Frequencies, H. B. Lee . . . 447

A Minimal Representation for a Cascade of Lattice Networks, D. M. Layton and I. T. Frisch 467

Analysis of Network by Vertex Reduction, H. Frank 485

Circuit Transformation Based Upon Topology, W. W. Happ, R. D. Lamb, and E. R. Robbins . . . 508

BANQUET

Toastmaster: B. R. Myers University of Notre Dame

Education for Systems Engineering, R. E. Machol, University of Illinois

GRAPH THEORY 530

Chairman: S. L. Hakimi, Northwestern University

Connectivity of Directed Graphs, Y. C. Chen and O. Wing 530

Generation of Symmetric n-Ports, B. R. Myers . . . 544

Synthesis of a Class of Singular Impedance Matrices, F. T. Boesch and C. L. Suffel 559

Digital Computer Analysis of Passive Networks Using Topological Formula, G. W. Zobrist and G. V. Lago 573

OPTIMUM AND LEARNING SYSTEMS 596

Chairman: B. C. Kuo, University of Illinois

Learning in Automatic Systems, J. Peschon 596

On the Learning Behavior of Finite-State Systems in Random Environment, G. J. McMurtry and K. S. Fu 618

State Increment Dynamic Programming, R. E. Larson 643

Suboptimal Linear Filtering for Continuous Dynamic Processes, J. S. Meditch 666

Synthesis of Approximately Optimum Control Systems by Conventional Bode Amplitude Diagrams, A. P. Sage and R. B. Streets 682

CIRCUIT THEORY 699

Chairman: B. A. Shenoi, University of Minnesota

Certain Constraints in the Admittance Descriminants of Networks, H. T. Lee 699

On the Properties of Certain Polynomials Encountered with Cascade Networks, K. A. Pullen 734

The Perfectly Coupled and Shunt-Augmented T, A New and Very General Two-Port Model in Passive Network Realization, K. H. Haase 745

ACTIVE CIRCUIT THEORY 765

Chairman: D. A. Calahan, University of Illinois

Fundamental Gain-Bandwidth Limitations of Non-Inverting Parametric Up-Converters, W. H. Ku 765

A 3-Transistor Gyrator Circuit with Variable Gyration Resistance, B. A. Shenoi 782

The Use of Topological Formulas in Analysing the Transistor Models of Linvill, R. H. Beene, E. D. Merkl, and R. H. Seacat 802

Generalized Impedance Leveling in Network Synthesis, M. A. Murray Lasso 820

CODES AND SEQUENTIAL MACHINES 841

The Determination of the Minimal TANT Networks, A. R. Meo 841

Monotone Feedback Shift-Registers, J. L. Massey and R. -w. Liu 864

Generalized Procedures for the Determination of Generating Functions of Directed Graphs of Discrete Systems, C. V. Ramamoorthy 875

A Graph Theoretic Approach to Linear Codes, W. D. Frazer 888

Error Corrections and Circuits on the n-Cube, R. T. Chien, C. B. Freiman, and D. T. Tang 899

INDEX OF AUTHORS 913

SOME RESULTS IN STABILITY THEORY

Charles A. Desoer
Department of Electrical Engineering
Electronics Research Laboratory
University of California, Berkeley, California[*]

The first result we wish to establish is a generalization of the Nyquist criterion.

ASSUMPTIONS

Following Nyquist(1) we consider the linear time-invariant single-loop feedback system shown on Figure 1. It will be referred to as the closed-loop system. The block labeled k is a constant gain factor (i. e., independent of time and frequency); if its input is $\eta(t)$ its output is $k \quad \eta(t)$ where k is a fixed number. The block labeled G satisfies the following conditions.[**]

---

[*] The research herein was jointly supported by National Science Foundation under Grant GP-2413 and National Aeronautics and Space Administration under Grant NsG-354(S-1).

[**] The discussion of stability for the case where the transfer functions are not rational is far from trivial. Any reader who doubts this should consider the function defined for $t > 0$ by $e^t \sin(e^t)$ and note that its Laplace transform is analytic for all finite s. This example shows that the discussion of stability cannot be settled by "looking at the singularity that is the furthest to the right," which is a legitimate procedure with rational transfer functions.

(G.1) Its input-output relation relating the output y, the zero-input response z and the input ζ is

$$y(t) = z(t) + \int_0^t g(t - \tau)\zeta(\tau)d\tau \qquad \text{for all } t \geq 0. \qquad (1)$$

(G.2) For all initial states, the zero-input response z is bounded on $[0, \infty)$ and tends to zero as $t \to \infty$. Let $z_M = \sup_{t \geq 0} |z(t)|$.

(G.3) The unit impulse response g is given by

$$g(t) = r + g_1(t) \qquad \text{for } t \geq 0$$

where the constant r is ≥ 0 and g_1 is bounded on $[0, \infty)$, tends to zero as $t \to \infty$, and is an element of $L^1(0, \infty)$. We shall write $G(s) = (r/s) + G_1(s)$.

The conclusion is stated in the form of:

Theorem I: Suppose the linear time-invariant single-loop-feedback system shown on Figure 1 satisfies assumptions (G.1), (G.2) and (G.3) and that $k > 0$. If the modified Nyquist diagram* of $G(s)$ does not encircle nor go through the critical point $(-1/k, 0)$, then

(a) the impulse response of the closed-loop system is bounded, tends to zero as $t \to \infty$, and is an element of $L^1(0, \infty)$;

(b) for any initial state, the zero-input response of the closed-loop system is bounded and goes to zero as $t \to \infty$;

(c) for any initial state and for any bounded input, the response of the closed-loop system is bounded.

---

*The modified Nyquist diagram is the map under G of the imaginary axis from which the interval $[-j\epsilon, j\epsilon]$ has been removed and replaced by the semi-circle $\epsilon e^{j\theta} | -(\pi/2) \leq \theta \leq (\pi/2)$. Here ϵ taken arbitrarily small.

If the Nyquist diagram of $G(s)$ encircles the critical point $(-1/k, 0)$ a finite number of times then the impulse response of the closed-loop system grows exponentially as $t \to \infty$.

It should be stressed that the only assumption that is made concerning the box G is that it fulfills the conditions (G.1), (G.2), and (G.3). Such conditions are often fulfilled by the impulse response of systems described by ordinary differential equations, difference-differential equations, and those whose input-output relation is obtained through the solution of partial differential equations. The latter is the case for distributed circuits and for many control systems.

ANALYSIS

Let u be the bounded input applied to the system and let $u_M \overset{\Delta}{=} \sup_{t \geq 0} |u(t)|$. The response of the closed-loop system starting from an arbitrary initial state is given by

$$y(t) = z(t) + k \int_0^t g(t-\tau)\,[u(\tau) - y(\tau)]\, d\tau \quad \text{for all} \quad t \geq 0. \tag{2}$$

The outline of the proof is as follows: (a) Using the Gronwall-Bellman(2, 3) inequality we establish that the solution of (2) is of exponential order, hence that Laplace transform techniques are ~~applicable.~~ (b) Simple considerations(4, 5) establish the uniqueness of the solution. (c) h, the unit impulse response of the closed-loop system, is shown to be bounded, to tend to zero and $\in L^1(0, \infty)$. The fact that $h \in L^1(0, \infty)$ is obtained by the use of the Principle of the Argument(7) and from a theorem of Paley and Wiener.(8) Reference to the integral equation defining h shows easily that h is bounded and goes to zero. The latter requires that it be shown that h is uniformly continuous. (d) z_c, the zero-input response of the closed-loop system, is bounded and goes to zero. (e) To obtain the converse, (i.e., if the Nyquist diagram encircles the critical point then the closed-loop impulse response grows exponentially) we must use a theorem of Doetsch.(10)

CONCLUSION

If we collect that which is known about the Nyquist criterion we may state that its virtues are the following: (a) It is a test which utilizes only experimentally available data (results of sinusoidal steady state measurements). (b) It is a necessary and sufficient condition, i.e., it does not lead to conditions that are too restrictive, as is often the case with Lyapunov techniques. (c) It covers an extremely wide class of linear time-invariant systems. (d) It leads to very specific conclusions regarding the impulse response, zero-input response and the total response of the closed loop system.

The second result, closely related to the first, is an extension of the Popov criterion, in the principal case.(11) Aizerman and Grantmacher have written a detailed monograph on this subject with an extensive bibliography.(12) In the U.S. it has also received attention.(13, 14) The approach used in these presentations is to make very detailed and specific assumptions concerning the internal dynamics of the linear part of the system. On the other hand one of the most important and appealing features of the Popov criterion is that it shares all the desirable features of the Nyquist criterion; the essence of the Nyquist criterion is that in its general formulation, as given above, it does not make any assumptions concerning the internal dynamics but rather makes assumptions only on its input-output relation. We propose to do the same for the Popov criterion.

DESCRIPTION OF THE SYSTEM

Figure 2 shows the single-input single-output feedback system under consideration; N is a time-varying memoryless nonlinearity, G is a linear time-invariant subsystem. N is assumed to be characterized as follows:

(N.1) If, at time t, its input is $\eta(t)$, then its output is $\varphi(\eta(t), t)$, where, for each $t \geq 0$, φ is a continuous function of its first argument and there are two positive numbers ϵ and k such that

$$0 < \epsilon \leq \frac{\varphi(\xi, t)}{\xi} \leq k - \epsilon \qquad \text{for all } \xi \neq 0, \text{ for all } t \geq 0. \qquad (1')$$

G is assumed to be characterized as follows: if α is its input and z is its zero-input response, its output y is given by

$$y(t) = z(t) + \int_0^t g(t - \tau)\, \alpha(\tau)\, d\tau \qquad \text{for all } t \geq 0. \qquad (2')$$

In addition the following conditions are satisfied:

(G.1) for any initial state, the zero-input response z of G is (i) bounded on $[0, \infty)$, (ii) uniformly continuous on $[0, \infty)$, and (iii) both z and $\dot{z}$ are elements of $L^2(0, \infty)$. Let $z_M \triangleq \sup_{t \geq 0} |z(t)|$.

(G.2) g, the unit impulse response of G, is an element of $L^1(0, \infty)$.

It is very important to note that since g is a special instance of a zero-input response, (G.1) implies that (i) g is bounded, (ii) g is uniformly continuous and (iii) g and $\dot{g}$ are elements of $L^2(0, \infty)$.

By Lemma 1 below, (G. 2) and (ii) imply that $g(t) \to 0$ as $t \to \infty$. Similarly $z(t) \to 0$ as $t \to \infty$. For future reference, let

$$g_M \triangleq \sup_{t \geq 0} |g(t)| .$$

The main result may be stated in the form of:

Theorem II: Consider the single-input single-output nonlinear time-varying single-loop-feedback system shown on Figure 2 where N is a time-varying memoryless nonlinearity which satisfies condition (N. 1) and G be a linear time-invariant subsystem satisfying conditions (G. 1) and (G. 2). If there exists a real number $q \neq 0$ and a positive δ such that

$$\operatorname{Re}\left[(1 + qj\omega)G(j\omega) + \frac{1}{k}\right] \geq \delta > 0 \qquad \text{for all } \omega \geq 0 \tag{P_δ}$$

then, for any initial state, the zero-input response y is bounded on $[0, \infty)$ and tends to zero as $t \to \infty$.

The proof of the theorem is based on three lemmas which are easy to prove and the proofs are in the appendix.

Lemma 1: Let f map $[0, \infty)$ into the real line R and be uniformly continuous. Let g map R into R, be continuous; let $g(0) = 0$ and let $x \neq 0$ imply $g(x) > 0$. Under these conditions, if

$$\int_0^\infty g[f(t)]\, dt < \infty \tag{3}$$

then

$$\lim_{t \to \infty} f(t) = 0. \tag{4}$$

Lemma 2: Let f_1, f_2, f_3 map $[0, \infty)$ into R and belong $L^2[0, \infty)$. Call F_1, F_2, and F_3 their Fourier transforms. Let H be the Fourier transform of a real-valued function h.

If

$$F_1(j\omega) = -H(j\omega)F_3(j\omega) + F_2(j\omega) \qquad \text{for all } \omega \tag{5}$$

and if there is a number σ_0 such that

$$\mathrm{Re}\, H(j\omega) \geq \sigma_0 > 0 \qquad \text{for all } \omega \geq 0 \tag{6}$$

then

$$\int_0^{\infty} f_1(t)f_3(t)\, dt \leq \frac{1}{8\pi} \int_{-\infty}^{+\infty} \frac{|F_2(j\omega)|^2}{\mathrm{Re}\, H(j\omega)}\, d\omega \leq C \tag{7}$$

where

$$C \triangleq (8\pi\, \sigma_0)^{-1} \int_{-\infty}^{+\infty} |F_2(j\omega)|^2\, d\omega = (4\sigma_0)^{-1} \int_{-\infty}^{+\infty} |f_2(t)|^2 dt. \tag{8}$$

Comment: Let us interpret $H(j\omega)$ as the impedance of a linear time-invariant network which is passive in view of (6). Let f_3 be the current and f_2 be the internal Thevenin equivalent voltage source. (See Figure 3.) The product $f_1(t)f_3(t)$ is the power delivered at time t by the network to the "external world." Note that the "external world" is any circuit (not necessarily linear nor time-invariant), which to the current f_3 creates the voltage drop f_1. Thus (7) asserts that, whatever may be the external circuit to which the network is connected, the energy delivered by the network in the interval [0, ∞) is bounded as indicated by the inequality (8). Note also that the equation is satisfied with the equality sign if the network is terminated by a linear time-invariant circuit defined by its impedance $Z(j\omega) = H^*(j\omega) = \mathrm{Re}\, H(j\omega) - j\mathrm{Im}\, H(j\omega)$: indeed the current is then $F_3(j\omega) = F_2(j\omega)/2\mathrm{Re}\, H(j\omega)$ and the power delivered at the frequency ω is $|F_2(j\omega)|^2/[4\mathrm{Re}\, H(j\omega)]$.

Lemma 3: In the proof of the theorem, q may be assumed to be positive.

Proof of the theorem: Let M be the linear time-invariant system whose transfer function is for each real ω given by $(1 + qj\omega)\,G(j\omega) + 1/k$. According to Lemma 3 we may assume $q > 0$. If $M\alpha$ represents the zero-state response of M to α, we have

$$\eta + q\dot{\eta} - \frac{\alpha}{k} = -M\alpha - z - q\dot{z}\,.$$

Let η_T be the function defined by

$$\eta_T(t) = \begin{cases} \eta(t) & 0 \le t \le T \\ 0 & \text{elsewhere.} \end{cases}$$

Let α_T, z_T, $\dot{z}_T$ be similarly defined. Since

$$\dot{\eta}(t) = -\dot{z}(t) - g(0+)\alpha(t) - \int_0^t \dot{g}(t-\tau)\alpha(\tau)\,d\tau\,.$$

(G.1) and (G.2) imply that $z + q\dot{z}$ belongs to $L^2(0, \infty)$, and that α_T, $\eta_T + q\dot{\eta}_T - \alpha_T/k$ are in $L^2(0, T)$ for all positive T. These conditions together with P_δ imply, by Lemma 2, that

$$\int_0^T (\eta_T(t) + q\dot{\eta}_T(t) - \alpha_T(t)/k)\,\alpha_T(t)\,dt \le C \qquad \text{for all } T > 0$$

where the constant C is independent of T because $z + q\dot{z} \in L^2(0, \infty)$. Expressing α in terms of η we get

$$k^{-1}\int_0^T [\,k\eta(t) - \varphi(\eta(t), t)]\ \varphi(\eta(t), t)\,dt + q\int_0^T \varphi(\eta(t), t)\,\dot{\eta}(t)\,dt \le C. \tag{9}$$

Call J_1 the first term of the left-hand side. Define the function $\tilde{\eta}$ on $[-1, 0)$ to be $\eta(0)(1 + t)$ for $-1 \le t \le 0$. From (1) we obtain

$$\frac{\epsilon}{2}\eta(0)^2 \leq \int_{-1}^{0} \varphi(\tilde{\eta}(t), t)\, \tilde{\eta}(t)\, dt \leq \frac{k}{2}\, \eta(0)^2 .$$

Let us add this integral to both sides of (9), hence

$$J_1 + q\, \frac{\epsilon}{2}\, \eta(T)^2 \leq C + \frac{k}{2}\, \eta(0)^2 .$$

(1') implies that the integrand of J_1 is non-negative for all t. Since $q > 0$ we get

$$\eta(T) \leq \sqrt{\frac{2C}{\epsilon q} + \frac{k}{\epsilon q}\, \eta(0)^2} \qquad \text{for all } T > 0.$$

Since the right-hand side is independent of T, η is bounded on $[0, \infty)$. Let $\eta_M \triangleq \sup\limits_{t \geq 0} |\eta(t)|$. We now observe that η is uniformly continuous on $[0, \infty)$: from the defining relation of η we get

$$|\eta(t+\xi) - \eta(t)| \leq |z(t+\xi) - z(t)| + \int_0^t |g(t+\xi-\tau) - g(t-\tau)|\, |\alpha(\tau)|\, d\tau$$
$$+ \int_t^{t \pm \xi} |g(t+\xi-\tau)\, \alpha(\tau)|\, d\tau .$$

Using (N.1) to bound α, we get

$$\leq |z(t+\xi) - z(t)| + k\eta_M \int_0^{\infty} |g(t'+\xi) - g(t')|\, dt' + k\eta_M g_M \xi .$$

(G.1) implies that the first term goes to zero as $\xi \to 0$ uniformly in t; $g \in L^1(0, \infty)$ implies that the integral $\to 0$ as $\xi \to 0$, and the same obviously holds for the last term. Therefore η is uniformly continuous on $[0, \infty)$. Thus all assumptions of Lemma 1 apply to J_1 and hence $\eta \to 0$ as $t \to \infty$. This concludes the proof of the theorem

CONCLUSION

The conclusions that can be drawn in this case concern only the zero-input response. On the other hand the class of nonlinear systems covered by Theorem II is very broad.

APPENDIX

Proof of Lemma 1: Suppose f does not go to zero as $t \to \infty$, then given any $\epsilon > 0$, there is an infinite sequence $t_1, t_2, \ldots, t_k, \ldots$ of points going to infinity such that

$$|f(t_k)| > 2\epsilon \qquad \text{for } k = 1, 2, \ldots$$

Since f is uniformly continuous, given any $\epsilon > 0$, there is a δ such that $|\xi| \leq \delta$ implies that $|f(t+\xi) - f(t)| < \epsilon$. Therefore in each one of the intervals $[t_k - \delta, t_k + \delta]$, $|f(t)| > \epsilon$. Since there is an infinite number of them, the integral in (3) of Lemma 1 must be infinite, which contradicts the hypothesis. Hence $f(t) \to 0$ as $t \to \infty$.

Proof of Lemma 2: The proof is a straightforward application of the completing the square technique.$^{(12)}$

Proof of Lemma 3: Suppose the condition (P_δ) is satisfied by the given G for some q, k and δ. The equation of the system may be symbolically represented by

$$G\varphi(\eta, t) = -\eta$$

where G denotes the effect of the linear time-invariant system G upon $\varphi(\eta, t)$, which is the output of N. Let $\hat{\varphi}(\eta, t) \triangleq k\eta - \varphi(\eta, t)$. After a few algebraic manipulations we obtain

$$H\varphi(\eta, t) = -\eta$$

which is of the same form as the preceding relation and where H is the linear system whose transfer function is $-G(s)[1 + kG(s)]^{-1}$. First from the definition of $\hat{\varphi}$ and (1) it follows that

$$0 < \epsilon \leq \frac{\hat{\varphi}(\xi, t)}{\xi} \leq (k - \epsilon) \quad \text{for all } \xi \neq 0 \text{ and all } t \geq 0.$$

In other words, $\hat{\varphi}$ satisfies the same conditions as φ. Second, we must verify that H satisfies the conditions (G. 1) and (G. 2). The geometric interpretation of the condition $(P_\delta)^{(12)}$ which is satisfied by G implies that the Nyquist diagram of G does not encircle the critical point (-1/k, 0). It follows from the theory of the Nyquist criterion given above that h, the impulse response of $H(s) = G(s)[1 + k\,G(s)]^{-1}$, is bounded on $[0, \infty)$, belongs to $L^1(0, \infty)$ and $\to 0$ as $t \to \infty$. Since $h(t) = g(t) - k(h * g)(t)$ and $\dot{h}(t) = \dot{g}(t) - k\,g(0+)h(t) - k\,(h * \dot{g})(t)$, for any finite $T > 0$, $\dot{h}$ is in $L^2(0, T)$ because each term of the right-hand side of the last equation is in $L^2(0, T)$ for every $T > 0$. To test the conditions (G. 1) note that z_c, the closed-loop zero-input response satisfies the equation $z_c(t) = z(t) - k(h * z)(t)$. Thus, immediately z_c is bounded and is uniformly continuous on $[0, \infty)$. Also $\dot{z}_c(t) = \dot{z}(t) - k\,h(0+)\,z(t) - k(h * \dot{z})(t)$. From these two equations, it follows that both z_c and $\dot{z}_c$ are in $L^2(0, \infty)$ since each term of the right-hand side is in $L^2(0, \infty)$. Thus by a process of relabeling we have transformed the given system characterized by G and φ into a system characterized by H and φ, such that H and φ satisfy the conditions (N. 1), (G. 1), and (G. 2). Finally, a direct calculation establishes that, for all ω,

$$\operatorname{Re}\ (1 - qj\omega)\,H(j\omega) + \frac{1}{k} = \frac{1}{|1 + kG(j\omega)|^2}\ \operatorname{Re}\ (1 + qj\omega)G(j\omega) + \frac{1}{k} \quad .$$

Therefore if the condition (P_δ) is satisfied by G for some $q < 0$ and some k, the same condition P_δ (with a different but still positive δ) is satisfied by H, for $-q$ and the same k.

REFERENCES

1. H. Nyquist, "Regeneration Theory," Bell System Tech. J., Vol. 11 (Jan. 1932) pp. 126-147.
2. E. A. Goddington and N. Levinson, Theory of Ordinary Differential Equations, New York: McGraw-Hill Book Co., 1955. P. 37, Problem 1.
3. N. Bourbaki, Fonctions d'une Variable Réelle, Paris: Hermann, 1961. Chap. IV, Vol. 1, No. 4, Lemme 2.
4. G. Doetsch, Handbuch der Laplace Transformation, Vol. I, Basel: Verlag Birkhäuser, 1950. P. 33.
5. G. Doetsch, loc. cit. P. 72.
6. G. Doetsch, loc. cit. P. 162
7. E. Hille, Analytic Function Theory, Vol. I, Ginn and Company, 1959. P. 252.
8. R. E. A. C. Paley and N. Wiener, Fourier Transforms in the Complex Domain, New York: American Mathematical Society, 1934. P. 60-61.
9. R. R. Goldberg, Fourier Transforms, Cambridge: Cambridge University Press, 1961. P. 4.
10. G. Doetsch, loc. cit., Vol. II. P. 150.
11. V. M. Popov, "Absolute stability of nonlinear systems of automatic control," Avtomat. i Telemekh. Vol. 22, August, 1961. P. 961-979.

12. M. Aizerman and F. R. Gantmacher, Absolute Stability of Regulator Systems, (Translation from Russian by E. Polak). San Francisco: Holden Day, 1964.

13. R. W. Brockett, "On the stability of nonlinear feedback systems," 1964 JACC Proceedings. P. 288-294.

14. E. I. Jury and B. W. Lee, "On the stability of a certain class of nonlinear sampled-data systems," IEEE Trans. Vol. AC-9, No. 1, Jan. 1964. P. 51-61.

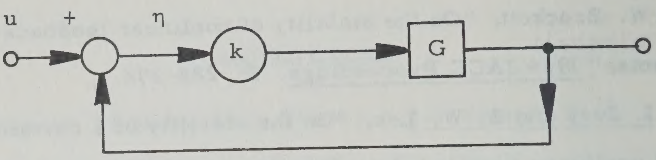


Figure 1

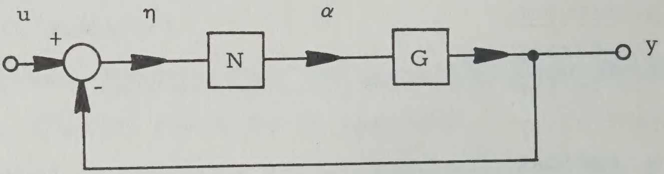


Figure 2

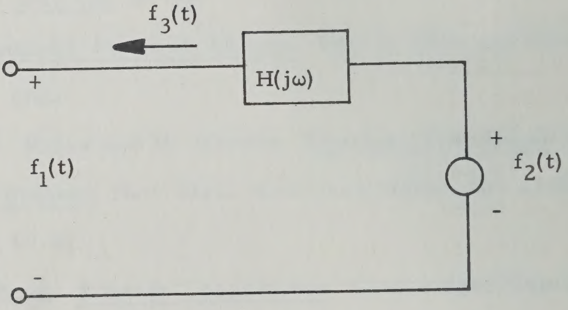

f_3(t)
H(jω)
+
f_1(t)
+
f_2(t)
-
-

Figure 3

DUAL DYNAMICAL SYSTEMS AND THEIR REPRESENTATION BY SYSTEM FUNCTIONS*

L. Weiss
Divisions of Applied Mathematics and Engineering
Brown University
Providence, R. I.

1. INTRODUCTION

The concept of the System Function has played a useful role in the input/output representation of linear differential systems. Since the early work of Zadeh [1] on the application of system functions to the analysis of linear time-varying systems, there have been numerous contributions to the subject which supplement and extend Zadeh's work [2], [3], [4], [5].

A concept which until now has been completely unrelated to the subject of system functions is that of dual dynamical systems which arises quite naturally (among other places) in the theory of controllability and observability of linear differential systems [6].

In this paper, we give an elementary exposition of the system function concept and then present some new results which show that there is an intimate connection between system functions and dual dynamical systems.

2. BACKGROUND

The purpose of this section is to define the class of systems to be considered and to present certain system theoretic concepts which are needed later on.

---

* This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research under Grant AF-AFOSR-693-64.

2.1 Systems to be Considered

We shall be concerned only with real, continuous-time, single-input, single-output linear systems whose input/output relation is described by an ordinary differential equation of finite order of the form

$$(1)\quad a_n(t)\frac{d^n y}{dt^n} + a_{n-1}(t)\frac{d^{n-1}y}{dt^{n-1}} + \dots + a_o(t)y$$

$$= b_m(t)\frac{d^m u}{dt^m} + \dots + b_o(t)u\ ,\ m < n$$

or

$$L\{y\} = M\{u\}$$

where

$$y = \text{output (scalar)}$$
$$u = \text{input (scalar)}$$

$$L\{\cdot\} = \sum_{i=o}^{n} a_i(t)\frac{d^i}{dt^i}\{\cdot\}$$

$$M\{\cdot\} = \sum_{i=o}^{m} b_i(t)\frac{d^i}{dt^i}\{\cdot\}$$

$$a_n(t) \neq 0 \text{ for any } t$$

and the values of m and n in (1) are minimal.

2.2 Causal and Anticausal Systems

The representation of a system by a differential equation (1) implies nothing about the ordering of the system's time scale [6]. Once the ordering is specified, the words "past" and "future" take on obvious meanings and provide the means for distinguishing between systems which obey the principle of causality and those which behave oppositely (the two cases of interest in this paper). More formally, we have:

DEFINITION 1. A causal system (C-system) is characterized by the property that all system responses at any time t_o are completely determined by past and present stimuli, i.e., inputs defined for $t \leq t_o$.

DEFINITION 2. An anticausal system (A-system) is characterized by the property that all system responses at any time t_o are completely determined by present and future stimuli, i.e., inputs defined for $t \geq t_o$.

The above concepts will play an important role in succeeding developments.

2.3 The Impulse Response of a System

It is well known that if (1) is a C-system, then the solution of (1) for zero initial conditions at initial time t_o can be put in the form

$$(2) \qquad y(t) = \int_{t_o}^{\infty} W_C(t,\tau)u(\tau)d\tau$$

where $W_C(t,\tau)$ is the causal (C-) impulse response of the system and has the property that

$$(3) \qquad W_C(t,\tau) = 0, \quad t < \tau .$$

The nomenclature for W_C is a result of the fact that one formally obtains $W_C(t,\tau)$ in (2) as the response at time t to an input u which is a unit impulse applied at time τ. The word "causal" follows from interpreting (3) as the mathematical expression of the statement "no response occurs prior to application of an input". Under certain conditions [6], the input/output relation (2) can be expressed with t_o replaced by $-\infty$, so that (2) becomes

$$(4) \qquad y(t) = \int_{-\infty}^{\infty} W_C(t,\tau)u(\tau)d\tau .$$

We shall assume that (4) is valid for all systems considered in this paper.

In like manner, if (1) is an A-system, then its solution for zero initial conditions at initial time t_o can be put in the form

$$(5) \qquad y(t) = \int_{-\infty}^{t_o} W_A(t,\tau)u(\tau)d\tau$$

where $W_A(t,\tau)$ is the anticausal (A) impulse response of the system and has the property that

$$(6) \qquad W_A(t,\tau) = 0, \quad t > \tau \ .$$

In this case, $W_A(t,\tau)$ is the response at time τ of a unit impulse applied at time τ but with respect to the previous time scale ordering we have the situation that the response precedes the application of an input.

Analogously to (4) we shall assume that the conditions hold [6] which allow us to express the input/output relation (5) in the form

$$(7) \qquad y(t) = \int_{-\infty}^{\infty} W_A(t,\tau)u(\tau)d\tau \ .$$

3. THE DEFINITION OF DUAL SYSTEMS

DEFINITION 3. Two systems S_1 and S_2, described by minimal order differential equations of the form (1) are said to be dual* systems with respect to their input/output relations if, for a given time scale ordering,

(i) S_1 is a C-system while S_2 is an A-system,
(ii) $W_1(t,\tau) = W_2(\tau,t)$

where

W_1 is the C-impulse response of S_1
W_2 is the A-impulse response of S_2 .

---

* The definition of dual systems given here essentially corresponds to the definition of adjoint systems given in [7].

It should be noted that the definition above is given only with respect to the input/output relations of two systems. There is a natural way of defining dual systems which takes into account their internal structure, i.e., their "state" representation. These systems are also dual in the sense of the above definition. The converse is not necessarily true, however [6].

4. REPRESENTATION OF SYSTEMS BY SYSTEM FUNCTIONS

In this section we define a system function and discuss its relevence to input/output calculations.

4.1 Definition of a System Function

We first define an elementary signal $k(t,\lambda)$ as being the kernel belonging to the inverse of an integral transform. Hence, by the nature of the operation of an integral transform, we know that there is a set of real and/or complex valued functions (generally forming a normed linear space) which can be expressed as the result of a linear operation on the elementary signal. For example, $\frac{1}{2\pi i} e^{\lambda t}$ is an elementary signal and all functions $f(\cdot)$ of exponential type and bounded variation can be expressed as

$$f(t) = \int_c F(\lambda) \frac{e^{\lambda t}}{2\pi i} d\lambda$$

where $F(\lambda)$ is defined by the relation

$$F(\lambda) = \int_{-\infty}^{\infty} f(t)e^{-\lambda t}dt$$

and c is a closed contour on the λ-plane which contains all the poles of $F(\lambda)$.

Now suppose we consider a system, under zero initial conditions, as an operator S which maps the system's input space [7] into its output space [7]. If the elementary signal k (a function of two variables, t and

λ) is in the domain of S, i.e., if $S\{k\}$ makes sense, then we define a system function H (which, in general, is a function of the same two variables t and λ) by the relation

$$S\{k\} = H\cdot k \, . \tag{8}$$

Remark: For convenience, the above statement may be roughly interpreted as defining a system function as a formal ratio of "output" to elementary signal "input." This has been the usual mode of expression in the past. It must be understood, however, that whenever this is done, no operation of division is necessarily implied and the ratio is simply a notation for the statement (8).

4.2 System Functions and Input/Output Representations

Based on the preceding statements, we have

Theorem 1. *Consider a system (1) with zero initial conditions. We define the following quantities:*

K = *integral operator which maps input space into output space of system.*

L = *a transform operator with continuous kernels.*

s = *elementary signal associated with transform operator* L.

U = *input function in the domain of* L.

P_U = *integral operator such that* $U = P_U\{s\}$.

Y = *output of system corresponding to input* U.

H = *system function based on* s.

Then, as long as the operators K *and* P_U *are commutative with respect to* s, *the output of the system for any allowable input* U *can be obtained using the system function* H.

Proof: Define y such that

$$y = K\{s\}.$$

Then, by definition

$$y = Hs.$$

Now,

$$U = P_U\{s\}.$$

We can therefore write

$$Y = K\{U\} = K\{P_U\{s\}\}.$$

If K and P_U commute with respect to s then

$$Y = P_U\{K\{s\}\} = P_U\{Hs\}. \tag{9}$$

Hence, given s and U, if H is known we can compute Y.

Q.E.D.

Remarks: (i) The existence of P_U is assured by the definition of U and s, and the construction of P_U follows immediately in any specific case by noting that

$$P_U\{s\} = L^{-1}\{LU\}. \tag{10}$$

The operator P_U will be uniquely defined for a given elementary signal s as long as the L transform of U is uniquely defined. In all cases of practical interest the transform is unique.

(ii) The assumption of commutativity of K and P_U with respect to s is simply another way of saying that one can reverse the order of integration when those two integral operators operate on s. The assumptions regarding the smoothness of the coefficients of (1) and the kernels of the operator L will, by Fubini's theorem [8], always guarantee that the reversal of order of integration is valid.

It follows from (9) that $P_U\{\cdot\}$ is simply the inverse L transform integral with integrand $(L\{U\})(\cdot)$. Therefore (10) can be expressed as

$$Y = L^{-1}\{HL\{U\}\}.$$

This establishes the fundamental link between system function representation of systems and integral transform techniques for analyzing input-output behavior of systems.

It is therefore useful at this stage of the development to present a brief account of pertinent aspects of the transform method. For a full and general discussion of the latter, see [9].

4.3 The Role of Compatible Integral Transforms in System Representation.

Suppose we have a set of (time-varying) systems each member of which has a time-domain input-output relation of the form

$$f(R)\{y\} = g(R)\{u\} \tag{11}$$

where R is a real linear differential operator which cannot be written as a function of an operator of lower order and f,g are real polynomial functions. R is then known as the canonical operator [9] for the set of systems. If there exists an integral transform which, for a given class of input functions, converts (11) into an algebraic equation in one variable, we say that the transform is compatible with the operator R [9].

In all that follows, the transform is assumed to be of conventional type. That is, if the transform operator is denoted by P and its inverse by P^{-1}, then for a suitably chosen time function $f(\cdot)$, we have the transform pair

$$f(t) = \int_{c_k} F(\lambda)k(t,\lambda)d\lambda = P^{-1}_{(\lambda;t)}\{F(\lambda)\} \tag{12}$$

$$F(\lambda) = \int_{c_\ell} f(t)\ell(\lambda,t)dt = P_{(t;\lambda)}\{f(t)\} \tag{13}$$

where

$k(t,\lambda)$ = inverse transform kernel
c_k = integration path associated with inverse transform

$\ell(\lambda,t)$ = transform kernel
c_ℓ = integration path associated with transform

and a subscript $(\alpha;\beta)$ on the P or P^{-1} symbols means "integration taken with respect to α, with β a transform parameter."

The following facts regarding the transform and inverse transform kernels can be demonstrated [9].

(i) The inverse transform kernel $k(t,\lambda)$ formally satisfies an equation

$$R\{k(t,\lambda)\} = q(\lambda)k(t,\lambda) \tag{14}$$

where, without loss of generality, we can take

$$q(\lambda) = \lambda. \tag{15}$$

(ii) The transform kernel $\ell(\lambda,t)$ formally satisfies the equation

$$R^*\{\ell(\lambda,t)\} = q(\lambda)\ell(\lambda,t)$$

where R* is the formal adjoint* of R and $q(\lambda) = \lambda$.

It follows from the preceding development that the transform P*, which is compatible with the operator R*, is related to the transform operator P according to the relations below.

---

* If R has the form

$$R\{\cdot\} = \sum_{i=o}^{n} a_i(t)\frac{d^i}{dt^i}\{\cdot\}$$

then the formal adjoint of R has the form

$$R^*\{\cdot\} = \sum_{i=o}^{n} (-1)^i \frac{d^i}{dt^i}\{a_i(t)\cdot\}$$

(16) inverse transform kernel for P^*: $k^*(t,\lambda)(=\ell(\lambda,t))$

(17) transform kernel for P^*: $\ell^*(\lambda,t)(=k(t,\lambda))$.

Hence, for a suitable function $f(\cdot)$, we have the transform pair

$$(18)\quad f(t) = P^{*\,-1}_{(\lambda;t)}\{F(\lambda)\} = \int_{c_{k^*}} F(\lambda)k^*(t,\lambda)d\lambda = \int_{c_{k^*}} F(\lambda)\ell(\lambda,t)d\lambda$$

$$(19)\quad F(\lambda) = P^*_{(t;\lambda)}\{f(t)\} = \int_{c_{\ell^*}} f(t)\ell^*(\lambda,t)dt = \int_{c_{\ell^*}} f(t)k(t,\lambda)dt.$$

Of course, if R is a self-adjoint operator, the kernels $k(t,\lambda)$ and $\ell(\lambda,t)$ will be identical in form.

It is easy to check that the ratio of the P-transformed output of (11) to P-transformed input qualifies as a system function. In this case, since the P-transform is compatible with the canonical operator of the system, the associated system function is a function only of the transform parameter. Under such conditions, the elementary signal $k(t,\lambda)$ is called a characteristic signal for the system. While it would be desirable to have at hand a compatible transform for analyzing a given set of systems, the process of deriving such a transform (assuming it exists) in a given case can be an extremely difficult problem [9]. The system function idea is an attempt to characterize input/output relations of systems by using *in*compatible transforms.

4.4 System Functions and Dual Systems.

In this section we present the basis for the main results of the paper. We consider, for a given C-system, two system functions associated with the kernels of the general P and P* transforms defined in (12, 13) and (18, 19) respectively. To the earlier hypotheses associated with these transforms, we add the assumption that the paths of integration C_{ℓ} and $C_{\ell *}$ are over the real line.

It is shown that each of the aforementioned system functions can be defined from a time-domain or transform-domain point of view and that the connection between the two equivalent definitions is provided by the concept of dual systems given earlier.

The following notation is used in the development.

S_C = given C-system

$S_{\tilde{C}}$ = the dual of S_C

$y_C(y_{\tilde{C}})$ = time-domain output of $S_C(S_{\tilde{C}})$

$Y_C(Y_{\tilde{C}})$ = (P or P*) transform-domain output of $S_C(S_{\tilde{C}})$

$W_C(W_{\tilde{C}})$ = impulse response of $S_C(S_{\tilde{C}})$.

Note: W_C is a C-impulse response, $W_{\tilde{C}}$ is an A-impulse response.

The following table gives the definitions of the two system functions to be discussed.

(i) The System Function $H_C(\lambda,t)$

(22)

TIME-DOMAIN DEFINITION	$\dfrac{\text{output of } S_C}{\text{input}}$	input=$k(t,\lambda)$
TRANSFORM-DOMAIN DEFINITION	$\dfrac{P^*\{\text{output of } S_{\tilde{C}}\}}{P^*\{\text{input}\}}$	$P^*\{\text{input}\}=\ell^*(\lambda,t)$

(ii) The System Function $H_{\tilde{C}}(\lambda,t)$

(23)

TIME-DOMAIN DEFINITION	$\left.\frac{\text{output of } S_{\tilde{C}}}{\text{input}}\right\vert_{\text{input}=k^*(t,\lambda)}$
TRANSFORM-DOMAIN DEFINITION	$\left.\frac{P\{\text{output of } S_{\tilde{C}}\}}{P\{\text{input}\}}\right\vert_{P\{\text{input}\}=\ell(\lambda,t)}$

The time-domain definition of the system function H_C was originally given by Zadeh [1] for the special case where P is the Laplace or Fourier transform; the transform domain definition of the system function $H_{\tilde{C}}$ was given independently by Gersho [2] and Bello [3] also for the special case where P is the Laplace or Fourier transform.

We now present

Assertion 1: The time-domain and transform-domain definitions of $H_C(\lambda,t)$ are equivalent.

Proof: The input/output relation for S_C is given by

$$y_C(t) = \int_{-\infty}^{\infty} W_C(t,\tau)u(\tau)d\tau \ .$$

Hence, from the time-domain definition of H_C, we have

$$(24)^* \qquad H_C(\lambda,t)k(t,\lambda) = \int_{-\infty}^{\infty} W_C(t,\tau)k(\tau,\lambda)d\tau \ .$$

But, according to the definition of dual systems, (24) can be rewritten as

$$(25) \qquad H_C(\lambda,t)k(t,\lambda) = \int_{-\infty}^{\infty} W_{\tilde{C}}(\tau,t)k(\tau,\lambda)d\tau$$

---

* It is tacitly assumed throughout that such expressions converge. For a given transform, this entails appropriate restrictions on the coefficient of (1).

where $W_{\tilde{C}}$ is the impulse response of $S_{\tilde{C}}$ (the dual of S_C). Now, from (17) and (19), the right side can be expressed as

$$(26) \qquad \int_{-\infty}^{\infty} W_{\tilde{C}}(\tau,t)\ell^*(\lambda,\tau)d\tau = P^*_{(\tau;\lambda)}\{W_{\tilde{C}}(\tau,t)\}.$$

And since

$$(27) \qquad \ell^*(\lambda,t) = \int_{-\infty}^{\infty} \delta(\tau-t)\ell^*(\lambda,\tau)d\tau = P^*_{(\tau;\lambda)}\{\delta(\tau-t)\}$$

then, from (25), (26), and (27) plus the fact that the response of $S_{\tilde{C}}$ to an impulse input $\delta(\tau-t)$ is the function $W_{\tilde{C}}(\tau,t)$ we have

$$H_C(\lambda,t) = \frac{P^*_{(\tau;\lambda)}\{W_{\tilde{C}}(\tau,t)\}}{P^*_{(\tau;\lambda)}\{\delta(\tau-t)\}} = \left.\frac{P^*\{\text{output of } S_{\tilde{C}}\}}{P^*\{\text{input}\}}\right|_{P^*\{\text{input}\}=\ell^*(\lambda,\tau)}$$

which is the transform definition of H_C in (22). Q.E.D.

Assertion 2: The time-domain and transform-domain definitions of $H_{\tilde{C}}(\lambda,t)$ are equivalent.

Proof: Is completely analogous to that for Assertion 1 and is left as an exercise. (Replace W_C by $W_{\tilde{C}}$, k by k*, ℓ^* by ℓ, P* by P.)

4.5 Application of System Functions to Input/Output Representation

The significance of the alternate definitions for H_C and $H_{\tilde{C}}$, and the role played therein by the concept of dual systems, is given by the following theorem, which is a specialization of Theorem 1.

THEOREM 3. (i) Consider a C-system S_C with an associated dual system $S_{\tilde{C}}$. For a given input, the system function H_C defined in (22) provides a means for representing:

(a) The time-domain output of S_C

(b) The P* transform-domain output of $S_{\tilde{C}}$.

(ii) Similarly, for a given input, the system function $H_{\tilde{C}}$ defined in (23) provides a means for representing:

(a') The time-domain output of $S_{\tilde{C}}$

(b') The P transform-domain output of S_C.

Proof: The proof for part (ii) of Theorem 3 is analogous to that for part (i) and both follow essentially the pattern of the proof of Theorem 1. We shall therefore prove part (i) only and merely quote the remaining results needed to prove the theorem.

Proof of part (i): (a) Consider the input/output relation for S_C

$$(28) \qquad y_C(t) = \int_{-\infty}^{\infty} W_C(t,\tau)u(\tau)d\tau \ .$$

As long as u is a P-transformable function, we can write

$$(29) \qquad u(\tau) = P^{-1}_{(\lambda;\tau)}\{U(\lambda)\} = \int_{c_k} U(\lambda)k(\tau,\lambda)d\lambda$$

where

$$U(\lambda) = P_{(t;\lambda)}\{u(t)\}.$$

Substituting (29) into (28)

$$y_C(t) = \int_{-\infty}^{\infty} W_C(t,\tau) \int_{c_k} U(\lambda)k(\tau,\lambda)d\lambda d\tau \ .$$

Assuming the order of integration can be reversed (see proof of Theorem 1) we obtain

$$y_C(t) = \int_{c_k} U(\lambda) \int_{-\infty}^{\infty} W_C(t,\tau)k(\tau,\lambda)d\tau d\lambda$$

and, by (24), this becomes

$$y_C(t) = \int_{c_k} U(\lambda)H_C(\lambda,t)k(t,\lambda)d\lambda$$

or equivalently

(30) $$y_C(t) = [P^{-1}_{(\lambda;t)}\{H_C(\lambda,\xi)U(\lambda)\}]_{\xi=t}.$$

Thus, given an input u to the system S_C, the system function H_C allows one to obtain the time domain output of S_C.

(b) Let $y_{\tilde{C}}$ be the time-domain output for the system $S_{\tilde{C}}$, and $Y_{\tilde{C}}$ the P transform-domain output. Then

$$Y_{\tilde{C}}(\lambda) = P_{(t;\lambda)}\{y_{\tilde{C}}(t)\} = \int_{-\infty}^{\infty} y_{\tilde{C}}(t)\ell^*(\lambda,t)dt.$$

From (28) with W_C replaced by $W_{\tilde{C}}$ we have

$$Y_{\tilde{C}}(\lambda) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_{\tilde{C}}(t,\tau)u(\tau)d\tau\ell^*(\lambda,t)dt.$$

Assuming once again that the order of integration can be reversed yields

(31) $$Y_{\tilde{C}}(\lambda) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_{\tilde{C}}(t,\tau)\ell^*(\lambda,t)dtu(\tau)d\tau.$$

But from the transform domain definition of H_C (see (25) with k replaced by ℓ^*)

(32) $$H_{\tilde{C}}(\lambda,\tau)\ell^*(\lambda,\tau) = \int_{-\infty}^{\infty} W_{\tilde{C}}(t,\tau)\ell^*(\lambda,t)dt.$$

Substitution of (32) into (31) gives the result

$$Y_{\tilde{C}}(\lambda) = \int_{-\infty}^{\infty} H_C(\lambda,\tau)\ell^*(\lambda,\tau)u(\tau)d\tau$$

or equivalently,

(33) $$Y_{\tilde{C}}(\lambda) = [\, P^{*}_{(t;\lambda)}\{H_C(\lambda',t)u(t)\}\,]_{\lambda'=\lambda} \,.$$

Hence, $H_C(\lambda,t)$ provides a means (equation (33)) for obtaining the P*-transformed output of S_C when the input u is given.

In like manner, it is easily shown that if

$$U^*(\lambda) = P^{*}_{(t;\lambda)}\{u(t)\}$$

then

(34) $$y_{\tilde{C}}(t) = [\, P^{*-1}_{(\lambda;t)}\{H_{\tilde{C}}(\lambda,\xi)U^*(\lambda)\}\,]_{\xi=t}$$

(35) $$Y_C(\lambda) = [\, P_{(t;\lambda)}\{H_{\tilde{C}}(\lambda',t)u(t)\}\,]_{\lambda'=\lambda}$$

which provide the needed facts for proving part (ii) and hence the theorem.

REFERENCES

1. L. A. Zadeh, "Frequency Analysis of Variable Networks," Proc. IRE, Vol. 38, No. 2 (1950), pp. 291-299.
2. A. Gersho, "Characterization of Time-Varying Linear Systems," Proc. IEEE, Vol. 51, No. 1 (1963), p. 238.
3. P. Bello, "Time-Frequency Duality," IEEE Trans. on Information Theory, Vol. IT-10, No. 1 (1964).
4. W. Kaplan, Operational Methods for Linear Systems. Reading, Mass.: Addison-Wesley. 1962
5. J. B. Cruz, Jr. and M. E. Van Valkenburg, "The Synthesis of Models for Time-Varying Linear Systems," Proc. PIB Symp. on Active Networks and Feedback Systems (April 1960).
6. L. Weiss and R. E. Kalman, "Contributions to Linear System Theory," (to appear in the International Journal of Engineering Science. Also available as RIAS Technical Report 64-9).
7. L. A. Zadeh and C. A. Desoer, Linear System Theory. New York: McGraw Hill. 1963.
8. M. E. Munroe, Introduction to Measure and Integration. Reading, Mass.: Addison-Wesley. 1953.
9. L. Weiss, "On the Foundations of Transfer Function Analysis," (to appear in the International Journal of Engineering Science).

ON CANONICAL REALIZATIONS*

R. E. Kalman**
Stanford University
Stanford, California

I. BACKGROUND

A dynamical system may be described in two different ways:

(i) axiomatically (state-variable equations) or by

(ii) input-output relations ("black box" approach).

Theoreticians prefer the first method and experimentalists insist that only the second has meaning in the "real world". There arises the mathematical problem of reconciling the two.
One wants to know just how much external measurements can say about the internal workings of a system.

A rather complete theory of this problem was developed recently by the writer in the most common special case of finite-dimensional, linear, time-invariant systems. In part, this is a generalization of the system-theoretic work of Elmer Gilbert, Brockway McMillan as well as of R. J. Duffin and D. Hazony. Mathematically speaking, the problem turns out to an unusual generalization of classical linear algebra, especially the theory of elementary divisors and invariant factors. Full details may be found in References 1 & 2 .

The purpose of this note is to present a digest of the result in a way somewhat different from the more formal expositions to be published elsewhere.

---

*This research was supported in part by the United States Air Force under Contract No. AF 49(638)1206.

**This paper was presented by Major B. Morgan.

2. CONCEPT OF A REALIZATION

Let Z be a matrix of ratios of polynomials in the complex variable s with real or complex coefficients. We say simply that Z is a rational matrix.

We assume, for convenience[*], that $Z(\infty) = 0$ and call such matrices regular.

A realization of a regular rational matrix is a triple of constant matrices $\{F, G, H\}$ such that

$$H(sI - F)^{-1}G = Z(s) \qquad (I = \text{unit matrix}). \tag{2.1}$$

We may regard Z as specifying the input/output relations of a (finite-dimensional, linear, time-invariant) system, that is, Z is the type (ii) system description.

If we have a "realization", then we can write down a type (i) description as

---

[*] It is a minor matter to remove this restriction when the more general case is of interest. See Reference 1.

(2.2) $dx/dt = Fx + Gu(t)$, $y(t) = Hx(t)$,

where $u(t)$, $y(t)$ and $x(t)$ are the input, output and state (vectors) respectively. Now (2.2) may be "physically realized" to a high accuracy (e.g. by an analog computer). In other words, having determined $\{F, G, H\}$ we possess a mathematical prescription for a concrete physical system which will correspond to the abstract input-output relation Z.

Similar comments hold in other situations. Z may be interpreted as the transfer-function matrix of a sampled-data system, of a linear counting circuit, etc.

Realizations always exist. An irreducible realization is one for which the dimension of the square matrix F in (2.1) is minimal; then we write $n_{\min}(Z)$. Irreducible realizations can be proved to be isomorphic, i.e., any two sets of matrices $\{F_1, G_1, H_1\}$ and $\{F_2, G_2, H_2\}$ which satisfy (irreducibly) (2.1) are but two different matrix representations of the same abstract linear transformations in (2.2). In other words, the only arbitrariness in an irreducible realization corresponding to a fixed Z is the choice of the basis for the state vector x in (2.2). This is as it should be, since the state vector is an internal variable which, by definition, can never be identified directly from input/output data. Relabeling the states must have no effect on the input-output relations.

3. CONSTRUCTION OF AN IRREDUCIBLE REALIZATION

As is well known, irreducible realizations may be obtained by reducing an arbitrary realization.(3) This method is practical and concrete. For theoretical purposes, however, we need a direct way of constructing an irreducible realization. This problem leads to a generalization of classical linear algebra. We will see that the elementary divisors of the F matrix of an irreducible realization may be determined from the structure of the transfer function matrix Z. The corresponding classical result is that these elementary divisors can be determined from the polynomial matrix $(sI - F)$.

4. THE SMITH CANONICAL FORM

Every polynomial matrix P has the Smith canonical form

$$(4.1) \qquad P(s) = A(s)\Gamma(s)B(s)$$

where A, B are polynomial matrices with constant, nonzero determinant;

$$\Gamma(s) = \operatorname{diag}\,[\gamma_1(s), \gamma_2(s), \ldots],$$

the γ_k are polynomials, and γ_k divides γ_{k+1}. The γ_k are unique but the A, B are not. Factoring each γ_k into linear factors (over the complex field!) we obtain

$$(4.2) \qquad \gamma_k(s) = \prod_{\mu=1}^{q} (s - \lambda_\mu)^{q_{\mu k}};$$

the $(s - \lambda_\mu)^{q_{\mu k}}$ are known as the invariant factors of P.

The main use of this canonical form is to prove the following classical theorem, mentioned above:

The elementary divisors of a constant square matrix F are equal to the invariant factors of the polynomial matrix $P(s) = sI - F$.

5. THE SMITH-McMILLAN FORM

It is natural to try to generalize this result in such a way that the polynomial matrix P is replaced by the rational matrix $Z(s) = (sI - F)^{-1}$. This suggests how to define the invariant factors of a rational matrix. (Note that these factors are not simply the reciprocals of the invariant factors of Z^{-1}, if Z^{-1} happens to be a polynomial!). Guided by such or other intuition, the invariant factors of Z were first defined by McMillan in 1952, as follows:

Let $\psi(s)$ be the least common denominator of Z. Then ψZ is a polynomial matrix. Take a Smith canonical form $\psi Z = A\Gamma B$. Write

$$(5.1) \qquad \Gamma/\psi = E\Psi^{-1}, \quad E, \Psi \text{ diagonal.}$$

Require further that corresponding elements ε_k and ψ_k of E and Ψ may be relatively prime.

Then it turns out that the ε_k and ψ_k are independent of the choice of A, B; i.e., they are truly invariants of Z. Further $\psi_1 = \psi$, ψ_{k+1} divides ψ_k, and $\psi_k = 0$ if $k \geq r(Z)$. We factor ψ_k as in (4.2); this defines the *invariant factors* of Z.

With these definitions, we find that the anticipated result:

> (*the elementary divisors of a constant square matrix* F *are equal to the invariant factors of the rational matrix* $Z(s) = (sI - F)^{-1}$)

is indeed true.

A further set of invariant factors may be obtained by factoring the ϵ_k. These will play no role in the theory, however; in fact, their significance is quite mysterious at present.

The main result may be regarded as a generalization of the preceding theorem. We want to show that:

> *The elementary divisors of a constant square matrix* F *belonging to an irreducible realization of a regular rational matrix* Z *are equal to the invariant factors of* Z.

6. CANONICAL REALIZATIONS

A realization of Z may be constructed directly from its Smith-McMillan canonical form

$$Z = AE\Psi^{-1}B = C\Psi^{-1}B. \tag{6.1}$$

The matrix F is exhibited in the Jordan canonical form, according to the invariant factors of Z; the matrices H and G correspond explicitly (no calculations involved!) to coefficients of polynomials in C and B. The detailed formulae for $\{F, G, H\}$ are far too complicated to be described here; the reader is referred to References 1 and 2.

The important fact is that these realizations are irreducible. The proof follows essentially from properties of the Smith-McMillan form. In other words, the Smith-McMillan form is the proper kind of matrix factorization which leads to irreducible solutions of the factorization problem (2.1). More generally, it is shown (Reference 2) that corresponding to every classical canonical form of the square matrix F there is a special irreducible realization.

The proof of irreducibility also makes use of the fact that realizations are irreducible if and only if they are completely controllable and completely observable. The algebraic technique developed for this purpose allows one to generalize some of the "classical" lemmas(3) guaranteeing complete controllability and complete observability.

The preceding results imply not only the theorem stated at the end of the last section, but they also contain, as a very special case, the relation

$$(6.2) \qquad n_{\min}(Z) = \sum_k \delta(\psi_k(Z)).$$

Here δ( ) denotes the degree of a polynomial. This is a "closed form" expression for the smallest number of state variables needed to realize the input-output relation represented by Z. Formula (6.2) includes an earlier result of Elmer Gilbert (see Reference 3 for a detailed discussion).

7. McMILLAN'S DEGREE

The first definition of the degree of a rational matrix was given by McMillan. In the regular case, he defines $\delta_M(Z)$ to be the right-hand side of (6.2). We have then the theorem

$$(7.1) \qquad n_{\min}(Z) = \delta_M(Z).$$

McMillan has shown that $\delta_M(Z)$ is the minimum number of reactive elements necessary to realize a square matrix Z as the impedance function of a passive N-port. (The same number is sufficient if Z is positive real and if ideal transformers are admitted.) Relation (7.1) shows that $\delta_M(Z)$ is the absolute minimum number of energy storage elements, even if we admit active circuit components.

8. DUFFIN-HAZONY DEGREE

An equivalent definition of the degree, due to Duffin and Hazony$^{(4)}$ is the following:

$$\delta_{DH}(Z) = \max_A \delta(\pi_A(Z)),$$

where

$$\det (A + Z) = \pi_A(Z)/\rho_A(Z)$$

and π and ρ are relatively prime polynomials.

The Duffin-Hazony degree was introduced to simplify McMillan's exposition and to provide practical methods for calculating the degree. (In view of (7.1), any method which computes an irreducible realization of Z provides a forteriori also the degree of Z.) The ideas of Duffin and Hazony are closely related to those of McMillan. No proof of the equivalence of δ_{DH} and δ_M is given in 4 , but such a proof follows immediately from Corollary 2 of Duffin and Hazony:

$$\delta_{DH}(Z) = \sum_\mu h_\mu$$

where h_μ is the largest degree to which the μ-th pole of Z occurs in the determinant or in a minor of Z.

To prove this result, we note two trivial lemmata (see Reference 1):

A. The least common denominator of the $k \times k$ minors of Z is given by $\psi_1 \cdots \psi_k$.

B. If $Z = \sum_\mu Z_\mu$ (where each Z_μ has exactly one pole $(\neq \infty)$) and if the poles of the Z_μ are distinct, then

$$\psi_k(Z) = \prod_{\mu} \psi_k(Z_\mu).$$

Now let us calculate:

$$\delta_M(z) \overset{\Delta}{=} \sum_k \delta(\psi_k(Z)) = \sum_{\mu,k} \delta(\psi_k(Z_\mu)) = \sum_{\mu} h_\mu \overset{\Delta}{=} \delta_{DH}(Z).$$

The second equality follows from Lemma B. The third follows from the definition of the ψ_k and Lemma A, which show that

$$\sum_k \delta\big(\psi_k(Z_\mu)\big) = h_\mu.$$

This establishes the equivalence of δ_{DH} and δ_M for regular matrices. The result is then easily extended to the general case.(1)

9. "DYNAMICAL" PROOFS OF THE McMILLAN-DUFFIN-HAZONY THEOREMS

It is interesting to establish some properties of the degree (found earlier by McMillan and Duffin and Hazony) by the method of irreducible realizations. This brings out the "dynamical" content of the theorems.

I. $\delta(Z_1 + Z_2) \leq \delta(Z_1) + \delta(Z_2)$.

Proof. We take an irreducible realization for Z_1 as well as one for Z_2. The direct sum (in the obvious sense) of these realizations will be a realization of $Z_1 + Z_2$. But the latter realization may be reducible; in that case the inequality sign will hold. Q.E.D.

II. $\delta(Z_1Z_2) \leq \delta(Z_1) + \delta(Z_2)$.

Proof. We can realize Z_1Z_2 as a cascade connection (obvious sense) of irreducible realizations of Z_1 and Z_2. As before, we get a strict inequality when the compound realization is reducible. Q.E.D.

III. *If* Z^* *is a submatrix of* Z, *then* $\delta(Z^*) \leq \delta(Z)$.

Proof. We can realize the submatrix Z^* by deleting certain inputs and outputs in the dynamical system (2.2). This may be expressed mathematically by writing $Z^* = AZB$, where A and B are matrices whose elements are zeros and ones. The degree of a constant matrix is of course 0. Hence III is a special case of II. Q.E.D.

More generally, if $Z^* = AZB$ (where A, B are nonsingular matrices), we might say that Z^* represents an input-output relation obtained from Z via memoryless codes. Thus memoryless coding can never increase the minimal number of state variables required for a realization.

10. GENERAL CANONICAL FORM

Are *all* canonical realizations obtainable from the Smith-McMillan form? This question makes sense only if F is made unique by putting it in some canonical form, say, the Jordan form. Even then there exist realizations which cannot be deduced from the Smith-McMillan form. The difficulty is that the Smith-McMillan form has too many invariants; the invariance of the ϵ_k is not needed in realization theory. The obvious remedy is to try to find a more general canonical form.

The following canonical representation is developed in Reference 2.

Let $\mathcal{P}$ be the operator which sends the rational function $f(x) = g(s)/h(s)$ (g, h = polynomials) into the rational function $f = g(\mathrm{mod}\ h)/h$. Then we have:

$$(10.1) \qquad Z = \mathcal{P}[Q\Psi^{-1}P]$$

where Ψ is a diagonal matrix (as in the Smith-McMillan canonical form), while P, Q are polynomial matrices with the property that their $k \times k$ principal minors have no factors in common with ψ_k, $k = 1, \ldots, r$.

With this representation, G determines P and H determines Q in a 1-1 fashion; as before ψ determines F in a Jordan canonical form.

In other words, there is a 1-1 correspondence between irreducible realizations and the representation (10.1).

11. REFERENCES

1 R. E. Kalman, "Irreducible realizations and the degree of a rational matrix", to appear in SIAM Journal.

2 R. E. Kalman, "Canonical realizations", to appear.

3 R. E. Kalman, "Mathematical description of linear dynamical systems", SIAM J. Control, 1 (1963) 152-192.

4 R. J. Duffin and D. Hazony, "The degree of a rational matrix function", SIAM Journal, 11 (1963) 645-658.

DISCUSSION

L. Weinberg
Department of Electrical Engineering
University of Michigan
Ann Arbor, Michigan

I think that Kalman wanted to do more than just represent the impedance matrix in the form you have on the board that uses the state-space matrices F, G, and H. In the paper he presented at Allerton last year, "On a New Characterization of Linear Passive Systems", he stated he had a complete characterization of lossless n-ports without ideal transformers or mutual inductance; that is, passive, lossless reciprocal networks containing only L's and C's. This is an important unsolved problem.

In the paper, Kalman gives the necessary and sufficient conditions for the two-port. The conditions are simple and are given in terms of the residues at each pole. He states:

a) $k_{12} = 0$, b) $k_{11} - k_{12}$, or c) $k_{12} = k_{22}$.

However, there is one trouble with this: the conditions are incorrect. They correspond to uncoupled (i.e., joined by no branch) input and output driving-point functions or to an L network. Surely an L network is not the most general two-port. A simple counterexample is a T network of three capacitances.

Thus the conditions are not necessary.

R. E. Kalman:

I regret that I couldn't be present to answer Weinberg on the spot. His note is the kind of misunderstanding which is easily clarified orally but sounds silly and downright rude in writing.

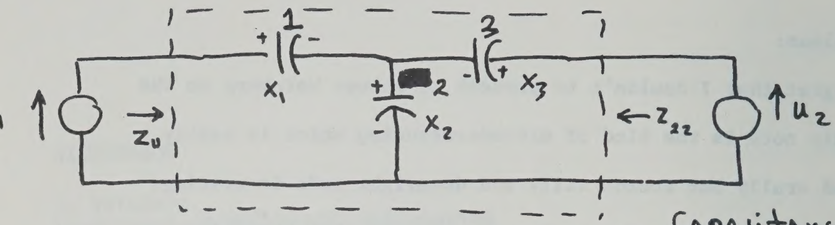
Of course, it is easier to become confused than to explain what the confusion is all about.

My paper at the Second Allerton Conference contains a lot more than just a method of representing an impedance matrix as $Z(s) = H(sI-F)^{-1}G$. The emphasis is on irreducible representations. The present paper has nothing to do with my Allerton paper of last year. This should have been clear to Weinberg, because I sent him a copy of the detailed version of the present paper already in June.

I am at a loss to explain how Weinberg formed the impression that I have "a complete characterization of lossless n-ports without ideal transformers or mutual inductance". I don't believe I have made precisely such a statement either in talking or in writing. I do recall having emphasized on several occasions that this is (as Weinberg says) an important and unsolved problem.

In the second paragraph Weinberg refers to a theorem stated without proof on p. 469 in the Proceedings of the First Allerton Conference. Here the question concerns 2-ports, not n-ports. Weinberg is worried because my result is "obviously" contradicted as a counterexample which even an undergraduate student should know.

His "counterexample" is of the following type:



Capacitances in farad

Fig. 1.

This network has the impedance matrix

$$Z(s) = \begin{bmatrix} \frac{3}{2s} & \frac{1}{2s} \\ \frac{1}{2s} & \frac{5}{6s} \end{bmatrix}$$

I will now give two independent arguments as to why this is not a counterexample to my result.

Weinberg's misunderstanding is of course caused by his overlooking the most important fact: my conclusions include the additional requirement that the network contain "a minimal number of reactive elements". Let us check whether or not his network has a minimal number of capacitors.

(1) (Arguments based on my first paper.) The system matrices describing the network of Fig. 1 (notations as in my first Allerton paper) are

$$A = 0, \qquad P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

It is easy to verify that this system is neither completely controllable nor completely observable. Hence Fig. 1 is not an irreducible realization of Z(·). Such cases are outside the scope of either of my Allerton papers.

(2) (Arguments based on my second paper.) A Smith-McMillan canonical form of Z(s) is given by

$$Z(s) = A(s)\,E(s)\,\Psi^{-1}(s)\,B(s),$$

$$= \begin{bmatrix} 9 & 3 \\ 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1/6 & 0 \\ 0 & 1/6 \end{bmatrix} .$$

Hence $\delta(Z) = 2$. The minimal number of reactive elements is therefore 2. Weinberg's network contains three capacitors. Hence it is not a minimal realization.

The whole emphasis in both of my Allerton papers has been on networks having a minimal number $(n = \delta(Z))$ reactive elements. Weinberg (and other classical network theorists) tend to overlook the significance of the minimal case. We must have a good theory for such networks because otherwise we cannot even hope to explain the phenomenon represented by Weinberg's hapless "counterexample": nonminimal realizations can sometimes be obtained using fewer element types than minimal realizations. (The Bott-Duffin procedure is the most famous known example of this sort.)

THE USE OF LUR'E FORMS TO ESTABLISH A SUFFICIENT CONDITION FOR STABILITY OF A CLASS OF DISCRETE FEEDBACK SYSTEMS WITH PARALLEL NONLINEAR ELEMENTS

G. J. Murphy and S. H. Wu

Information - Processing and Control Systems Laboratory

Northwestern University

Evanston, Illinois

INTRODUCTION

There has recently been much interest in the use of Lur'e forms in studying the stability of control systems.

The problem of determining the stability of a sampled-data feedback system containing a single nonlinear gain element has received much attention in recent years. Using a quadratic form of Liapunov function, Kodama(1) extended Aizerman's method so as to apply it to such systems. Pearson and Gibson(2) developed a method for constructing a Lur'e form(3,4) of Liapunov function for such systems, and Tsypkin,(5) Jury and Lee,(6) Kalman and Szego,(7) and Szego and Pearson(8) have studied the problem by use of Popov's method.(4) The investigation of the stability of a feedback control system containing several nonlinear gain elements connected in cascade has been treated by Ibrahim and Rekasius.(9) However, the problem of investigating the stability of sampled-data feedback systems containing several nonlinear gain elements connected in parallel, as shown in Fig. 1, has received virtually no attention.

The purpose of this paper is to present a method, based on a Lur'e form of Liapunov function, for establishing a sufficient condition for asymptotic stability in the large of systems of the kind illustrated in Fig. 1.

DEVELOPMENT OF THE MODEL

Let the transfer characteristics $v_i(e_1), i = 1,2,\ldots,N$, be such that

$$v_i(0) = 0, \; i = 1,2,\ldots,N \tag{1}$$

$$v_i(a) = -v_i(-a), 0<a<\infty, i = 1,2,\ldots,N \tag{2}$$

$$h_i(t) \triangleq \frac{\hat{v}_i(t)}{e_1(t)} > 0, e_1(t) \neq 0, i = 1,2,\ldots,N, \tag{3}$$

where $\hat{v}_i(t) \triangleq v_i[e_1(t)],$ (3a)

and

$$\frac{d\, v_i(e_1)}{d\, e_1} \leq \mu_i, -\infty<e_i<\infty, i = 1,2,\ldots,N, \tag{4}$$

where μ_i are finite, positive real constants. Then the input-output curves for the nonlinear elements are symmetric about the origin, pass through the origin, and have bounded slope. Typical characteristics are shown in Fig. 2.

If the i^{th} plant is of order n_i and if the state vector for the i^{th} plant is denoted by $\underset{\sim}{x}_i, i = 1,2,\ldots,N$, with elements $x_{i1}, x_{i2}, \ldots, x_{in_i}$, then the dynamic behavior of the plants is described by

$$\dot{\underset{\sim}{x}}_i(t) = \underset{\sim}{L}_i \underset{\sim}{x}_i(t) + \underset{\sim}{d}_i \hat{v}_i(jT), 0 \leq t-jT<T, j=0,1,2,\ldots, i=1,2,3,\ldots,N, \tag{5}$$

where $\underset{\sim}{L}_i$ is a matrix with n_i rows and n_i columns and $\underset{\sim}{d}_i$ is a vector with n_i elements (i.e., a matrix with n_i rows and 1 column). The solution to (5) is

$$\underset{\sim}{x}_i[(j+1)T] = \underset{\sim}{\Phi}_i(T)\underset{\sim}{x}_i(jT) + \underset{\sim}{u}_i(T)\hat{v}_i(jT),\ j=0,1,2,\ldots,i = 1,2,\ldots,N, \tag{6}$$

where

$$\underset{\sim}{\Phi}_i(T) \overset{\Delta}{=} \exp \underset{\sim}{L}_i T, i = 1,2,\ldots,N \tag{7}$$

and

$$\underset{\sim}{u}_i(T) \overset{\Delta}{=} \underset{\sim}{\Phi}_i(T)\int_0^T \underset{\sim}{\Phi}_i(-\tau)\, \underset{\sim}{d}_i d\tau, i = 1,2,\ldots,N \tag{8}$$

$$= \begin{bmatrix} u_{i1} & u_{i2} & \ldots & u_{in_i} \end{bmatrix}^t, i = 1,2,\ldots,N, \tag{9}$$

where the superscript t is used to denote the transpose.

If the state vector for the complete system is denoted by

$$\underset{\sim}{y}(t) \overset{\Delta}{=} \underset{\sim}{A} \begin{bmatrix} \underset{\sim}{x}_1(t) \\ \underset{\sim}{x}_2(t) \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{x}_N(t) \end{bmatrix} = [y_1(t)\ y_2(t)\ \ldots\ y_n(t)]^t, \tag{10}$$

where

$$n \overset{\Delta}{=} \sum_{i=1}^{N} n_i \tag{11a}$$

and where

$$\underset{\sim}{A} \overset{\Delta}{=} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \tag{11b}$$

is an arbitrary non-singular matrix, then it follows from (6) that

$$\underset{\sim}{y}[(j+1)T] = \underset{\sim}{\Psi}(T)\underset{\sim}{y}(jT) + \underset{\sim}{w}(jT), j = 0,1,2,\ldots, \tag{12}$$

where

$$\underset{\sim}{\Psi}(T) \triangleq \underset{\sim}{A} \begin{bmatrix} \underset{\sim}{\Phi}_1(T) & \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Phi}_2(T) & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{\Phi}_3(T) & \cdots & \underset{\sim}{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{\Phi}_N(T) \end{bmatrix} \underset{\sim}{A}^{-1} \qquad (13)$$

and

$$\underset{\sim}{w}(jT) \triangleq \underset{\sim}{A} \begin{bmatrix} \hat{v}_1(jT)\ \underset{\sim}{u}_1(T) \\ \hat{v}_2(jT)\ \underset{\sim}{u}_2(T) \\ \vdots \\ \hat{v}_n(jT)\ \underset{\sim}{u}_n(T) \end{bmatrix}, \ j = 0,1,2,\ldots. \qquad (14)$$

$$= \underset{\sim}{A} \ [\hat{v}_1(jT)u_{11}(T)\ \hat{v}_1(jT)u_{12}(T)\cdots\hat{v}_1(jT)u_{1n_1}(T)\cdots$$
$$\hat{v}_N(jT)u_{N1}(T)\hat{v}_N(jT)u_{N2}(T)\cdots\hat{v}_N(jT)u_{Nn_N}(T)]^t, j=0,1,2,\cdots. \qquad (15)$$

In order to study the stability of the system, let the system be unforced. That is, let

$$r(t) \equiv 0, \ 0 \leq t < \infty \ . \qquad (16)$$

In addition, let

$$c(jT) = \underset{\sim}{\alpha}^t \ \underset{\sim}{y}(jT), \ j = 0,1,2,\cdots, \qquad (17)$$

where

$$\underset{\sim}{\alpha} \triangleq [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_n]^t \ . \qquad (18)$$

Then

$$\hat{v}_i(jT) = -Kh_i(jT)\underset{\sim}{\alpha}^t\underset{\sim}{y}(jT), j=0,1,2,\ldots,i=1,2,\ldots N, \qquad (19)$$

where $h_i(jT)$ is defined as in (3).

Substituting (19) in (12) then gives

$$\underset{\sim}{y}[(j+1)T] = \underset{\sim}{\Psi}(T)\underset{\sim}{y}(jT) - K\ \underset{\sim}{D}[\underset{\sim}{y}(jT)]\underset{\sim}{y}(jT), j=0,1,2,\ldots, \qquad (20)$$

where

$$\underset{\sim}{D}[\underset{\sim}{y}(jT)] \triangleq \underset{\sim}{A} \begin{bmatrix} h_1(jT)u_{11}(T)\alpha_1 & h_1(jT)u_{11}(T)\alpha_2 & \cdots & h_1(jT)u_{11}(T)\alpha_n \\ h_1(jT)u_{12}(T)\alpha_1 & h_1(jT)u_{12}(T)\alpha_2 & \cdots & h_1(jT)u_{12}(T)\alpha_n \\ \vdots & \vdots & \cdots & \vdots \\ h_1(jT)u_{1n_1}(T)\alpha_1 & h_1(jT)u_{1n_1}(T)\alpha_2 & \cdots & h_1(jT)u_{1n_1}(T)\alpha_n \\ \vdots & \vdots & \cdots & \vdots \\ h_N(jT)u_{N1}(T)\alpha_1 & h_N(jT)u_{N1}(T)\alpha_2 & \cdots & h_N(jT)u_{N1}(T)\alpha_n \\ h_N(jT)u_{N2}(T)\alpha_1 & h_N(jT)u_{N2}(T)\alpha_2 & \cdots & h_N(jT)u_{N2}(T)\alpha_n \\ \vdots & \vdots & \cdots & \vdots \\ h_N(jT)u_{Nn_N}(T)\alpha_1 & h_N(jT)u_{Nn_N}(T)\alpha_2 & \cdots & h_N(jT)u_{Nn_N}(T)\alpha_n \end{bmatrix},$$

$$j = 0,1,2,\cdots. \qquad (21)$$

A reference condition is now established by letting

$$\underset{\sim}{D}[\underset{\sim}{y}(jT)] = \underset{\sim}{D}_0 + \underset{\sim}{D}_1[\underset{\sim}{y}(jT)],\ j=0,1,2,\cdots \qquad (22)$$

and

$$K = K_0 + \Delta K, \qquad (23)$$

where K_0 is a reference value of the gain constant K and $\underset{\sim}{D}_0$ is a reference matrix corresponding to some c(jT), for the nonlinear transition matrix $\underset{\sim}{D}(jT)$. Then (20) may be rewritten as

$$\underset{\sim}{y}[(j+1)T] = \underset{\sim}{B}(T)\underset{\sim}{y}(jT)-\underset{\sim}{M}[\underset{\sim}{y}(jT)]\underset{\sim}{y}(jT),\ j=0,1,2,\cdots, \qquad (24)$$

where

$$\underset{\sim}{B}(T) \triangleq \underset{\sim}{\Psi}(T) - K_0\underset{\sim}{D}_0 \qquad (25)$$

$$= \begin{bmatrix} b_{11}(T) b_{12}(T) & \cdots & b_{1n}(T) \\ b_{21}(T) b_{22}(T) & \cdots & b_{2n}(T) \\ \vdots \quad \vdots & \cdots & \vdots \\ b_{n1}(T) b_{n2}(T) & \cdots & b_{nn}(T) \end{bmatrix} \qquad (26)$$

and

$$\underset{\sim}{M}[\underset{\sim}{y}(jT)]\overset{\Delta}{=}K_0\underset{\sim}{D}_1[\underset{\sim}{y}(jT)] + \Delta\, K\, \underset{\sim}{D}[\underset{\sim}{y}(jT)],\ j=0,1,2,\cdots \tag{27}$$

$$= \begin{bmatrix} m_{11}[\underset{\sim}{y}(jT)] & m_{12}[\underset{\sim}{y}(jT)] & \cdots & m_{1n}[\underset{\sim}{y}(jT)] \\ m_{21}[\underset{\sim}{y}(jT)] & m_{22}[\underset{\sim}{y}(jT)] & \cdots & m_{2n}[\underset{\sim}{y}(jT)] \\ \vdots & \vdots & \cdots & \vdots \\ m_{n1}[\underset{\sim}{y}(jT)] & m_{n2}[\underset{\sim}{y}(jT)] & \cdots & m_{nn}[\underset{\sim}{y}(jT)] \end{bmatrix},\ j=0,1,2,\cdots. \tag{28}$$

CONSTRUCTION OF A LUR'E FORM OF LIAPUNOV FUNCTION

Let

$$V[\underset{\sim}{y}(jT)]\overset{\Delta}{=}\underset{\sim}{y}^t(jT)\underset{\sim}{P}\underset{\sim}{y}(jT) + \frac{\underset{\sim}{\delta}^t}{K^2}\int_0^{e_1(jT)} \underset{\sim}{v}(\xi)d\xi,\ j=0,1,2,\cdots, \tag{29}$$

where

$$\underset{\sim}{\delta} \overset{\Delta}{=} \begin{bmatrix} \delta_1 & \delta_2 & \cdots & \delta_n \end{bmatrix}^t \tag{30}$$

is independent of K and

$$\underset{\sim}{v}(\xi) \overset{\Delta}{=} \begin{bmatrix} v_1(\xi) & v_2(\xi) & \cdots & v_N(\xi) \end{bmatrix}^t \tag{31}$$

and the $v_i(\xi)$ are described by (1) through (4). Because of the conditions expressed in (1) and (3),

$$\int_0^{e_1(jT)} v_i(\xi)d\xi > 0,\ j = 0,1,2,\cdots. \tag{32}$$

It follows that if $\underset{\sim}{y}^t(jT)\underset{\sim}{P}\ \underset{\sim}{y}(jT)$ is positive definite and if each element of the vector $\underset{\sim}{\delta}$ is non-negative, then $V[\underset{\sim}{y}(jT)]$ is a positive definite scalar function. Let $\underset{\sim}{P}$ and $\underset{\sim}{\delta}$ now be so chosen.

It follows from (29) that

$$\Delta\, V[\underset{\sim}{y}(jT)] \overset{\Delta}{=} V\{\underset{\sim}{y}[(j+1)T]\}-V[\underset{\sim}{y}(jT)] = \underset{\sim}{y}^t[(j+1)T]\ \underset{\sim}{P}\ \underset{\sim}{y}[(j+1)T]$$
$$-\underset{\sim}{y}^t(jT)\underset{\sim}{P}\ \underset{\sim}{y}(jT) + \frac{\underset{\sim}{\delta}^t}{K^2}\int_{e_1(jT)}^{e_1[(j+1)T]}\underset{\sim}{v}(\xi)d\xi,\ j=0,1,2,\cdots \tag{33}$$

and it is a consequence of (4) that

$$\int_{e_1(jT)}^{e_1[(j+1)T]} v_i(\xi)d\xi \le v_i[e_1(jT)]\left\{e_1[(j+1)T] - e_1(jT)\right\} + \tfrac{1}{2}\,\mu_i\left\{e_1[(j+1)T] - e_1(jT)\right\}^2, j=0,1,2,\cdots. \tag{34}$$

Accordingly,

$$\begin{aligned}\Delta V[\underset{\sim}{y}(jT)] \le{}& \underset{\sim}{y}'[(j+1)T]\, \underset{\sim}{P}\, \underset{\sim}{y}[(j+1)T] - \underset{\sim}{y}^t(jT)\underset{\sim}{P}\, \underset{\sim}{y}(jT)\\ &+ \frac{\underset{\sim}{\delta}^t}{K^2}\Big\{\underset{\sim}{v}[e_1(jT)]\,[e_1(\overline{j+1}\ T) - e_1(jT)]\\ &+ \tfrac{1}{2}\,\underset{\sim}{\mu}\,[e_1(\overline{j+1}\ T) - e_1(jT))]^2\Big\}, j=0,1,2,\cdots,\end{aligned} \tag{35}$$

where

$$\underset{\sim}{\mu} \triangleq \left[\mu_1 \ \ \mu_2 \cdot \cdot \cdot \mu_N\right]^t. \tag{36}$$

Substitution of (24) in (35) and collection of terms yield

$$\Delta\ V[\underset{\sim}{y}(jT)] \le -\underset{\sim}{y}^t(jT)\underset{\sim}{R}[T,\underset{\sim}{y}(jT)]\underset{\sim}{y}(jT),\ j = 0,1,2,\cdots, \tag{37}$$

where

$$\begin{aligned}\underset{\sim}{R}[T,\underset{\sim}{y}(jT)] \triangleq{}& -\underset{\sim}{B}^t(T)\underset{\sim}{P}\ \underset{\sim}{B}(T) + \underset{\sim}{P} - \tfrac{1}{2}\underset{\sim}{\delta}^t\underset{\sim}{\mu}[\underset{\sim}{B}^t(T)\underset{\sim}{A}_1\underset{\sim}{B}(T) + \underset{\sim}{A}_1]\\ &+ \left\{\underset{\sim}{B}^t(T)\underset{\sim}{P}\ \underset{\sim}{M}[\underset{\sim}{y}(jT)] + \underset{\sim}{M}^t[\underset{\sim}{y}(jT)]\underset{\sim}{P}\ \underset{\sim}{B}(T)\right\}\\ &+ \tfrac{1}{2}\ \underset{\sim}{\delta}^t\underset{\sim}{\mu}\ \left\{\underset{\sim}{B}^t(T)\underset{\sim}{A}_1\underset{\sim}{M}[\underset{\sim}{y}(jT)] + \underset{\sim}{M}^t[\underset{\sim}{y}(jT)]\ \underset{\sim}{A}_1\underset{\sim}{B}(T)\right\}\\ &+ \tfrac{1}{2}\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})\left[\underset{\sim}{B}^t(T)\underset{\sim}{A}_1 + \underset{\sim}{A}_1{}^t\ \underset{\sim}{B}(T)\right]\\ &+ \underset{\sim}{\delta}^t\underset{\sim}{h}\left\{\underset{\sim}{A}_1 + \underset{\sim}{M}^t[\underset{\sim}{y}(jT)]\underset{\sim}{A}_1\right\} - \tfrac{1}{2}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\ \underset{\sim}{M}^t[\underset{\sim}{y}(jT)]\underset{\sim}{A}_1\underset{\sim}{M}[\underset{\sim}{y}(jT)]\\ &- \underset{\sim}{\delta}^t\underset{\sim}{\mu}\ \underset{\sim}{M}^t[\underset{\sim}{y}(jT)]\underset{\sim}{A}_1 - \underset{\sim}{M}^t[\underset{\sim}{y}(jT)]\underset{\sim}{P}\ \underset{\sim}{M}[\underset{\sim}{y}(jT)],\ j=0,1,2,\cdots,\end{aligned} \tag{38}$$

where

$$\underset{\sim}{A}_1 \triangleq \begin{bmatrix} \alpha_1^2 & \alpha_1\alpha_2 & \cdots & \alpha_1\alpha_n\\ \alpha_2\alpha_1 & \alpha_2^2 & \cdots & \alpha_2\alpha_n\\ \vdots & \vdots & & \vdots\\ \alpha_n\alpha_1 & \alpha_n\alpha_2 & \cdots & \alpha_n^2 \end{bmatrix}. \tag{39}$$

To simplify the expression on the right-hand side of (38), now let

$$\underset{\sim}{Q}(T) \triangleq -\underset{\sim}{B}^t(T)\underset{\sim}{P}\,\underset{\sim}{B}(T) + \underset{\sim}{P} - \tfrac{1}{2}\,\underset{\sim}{\delta}^t\underset{\sim}{\mu}\left[\underset{\sim}{B}^t(T)\underset{\sim}{A}_1\underset{\sim}{B}(T) + \underset{\sim}{A}_1\right]. \qquad (40)$$

Moreover, let

$$\underset{\sim}{F}(T) \triangleq -\left[\underset{\sim}{B}^t(T)\underset{\sim}{P}\,\underset{\sim}{B}(T) - \underset{\sim}{P}\right] \qquad (41)$$

and

$$\underset{\sim}{N}(T) \triangleq \tfrac{1}{2}\,\underset{\sim}{\delta}^t\underset{\sim}{\mu}\left[\underset{\sim}{B}^t(T)\,\underset{\sim}{A}_1\underset{\sim}{B}(T) + \underset{\sim}{A}_1\right]. \qquad (42)$$

Then

$$\underset{\sim}{Q}(T) = \underset{\sim}{F}(T) - \underset{\sim}{N}(T). \qquad (43)$$

Since $\underset{\sim}{\delta}^t\underset{\sim}{\mu} > 0$, $\underset{\sim}{\chi}^t(jT)\underset{\sim}{N}(T)\underset{\sim}{\chi}(jT)$ is non-negative definite. Since

$$\underset{\sim}{F}(T) = \underset{\sim}{Q}(T) + \underset{\sim}{N}(T), \qquad (44)$$

it follows then that $\underset{\sim}{\chi}^t(jT)\underset{\sim}{F}(T)\underset{\sim}{\chi}(jT)$ is positive definite if $\underset{\sim}{\chi}^t(jT)\underset{\sim}{Q}(T)\underset{\sim}{\chi}(jT)$ is positive definite. From (43) it is found that

$$\underset{\sim}{F}(T) = -\left[\underset{\sim}{B}^t(T)\,\underset{\sim}{P}\,\underset{\sim}{B}(T) - \underset{\sim}{P}\right]. \qquad (45)$$

It is known(1) that an equation of the form of (45) has the property that a unique real, symmetric solution $\underset{\sim}{P}$ exists for a given real, symmetric $\underset{\sim}{F}(T)$ if and only if $\lambda_i[\underset{\sim}{B}(T)]\,\lambda_j[\underset{\sim}{B}(T)] \neq 1$, $i = 1,2,\cdots,n$, where $\lambda_i[\underset{\sim}{B}(T)]$ denotes the i^{th} eigenvalue of $\underset{\sim}{B}(T)$. It is known(10) also that if $\underset{\sim}{F}(T)$ is positive definite, $\underset{\sim}{P}$ is also positive definite if $|\lambda_i[\underset{\sim}{B}(T)]| < 1$, $i = 1,2,\cdots,n$. It follows that if the nonlinear system is linearized to obtain the linear system

$$\underset{\sim}{\chi}[(j+1)T] = \underset{\sim}{B}(T)\underset{\sim}{\chi}(jT), \qquad (46)$$

if this linear system is stable, and if a positive definite real symmetric matrix $\underset{\sim}{Q}(T)$ is chosen in accordance with (43), then there exists a positive definite real, symmetric matrix $\underset{\sim}{P}$ uniquely defined by (45), and the $V[\underset{\sim}{\chi}(jT)]$ defined by (29) is a Liapunov function for the linearized system.*

The actual nonlinear system, described by (24), is asymptotically stable in the large if, in addition, the matrix $\underset{\sim}{R}[T,\chi(jT)]$ defined in (38) is positive definite for $j = 0,1,2,\cdots$.

DETERMINATION OF A RANGE OF GAIN FOR ASYMPTOTIC STABILITY IN THE LARGE

Substitution of (40) in (38) gives

$$\begin{aligned}\underset{\sim}{R}[T,\chi(jT)] &= \underset{\sim}{Q}(T) + \left\{\underset{\sim}{B}^t(T)\underset{\sim}{P}\,\underset{\sim}{M}[\chi(jT)] + \underset{\sim}{M}^t[\chi(jT)]\underset{\sim}{P}\,\underset{\sim}{B}(T)\right\}\\ &+ \tfrac{1}{2}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\left\{\underset{\sim}{B}^t(T)\underset{\sim}{A}_1\underset{\sim}{M}[\chi(jT)] + \underset{\sim}{M}^t[\chi(jT)]\underset{\sim}{A}_1\underset{\sim}{B}(T)\right\}\\ &+ \tfrac{1}{2}\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})[\underset{\sim}{B}^t(T)\underset{\sim}{A}_1 + \underset{\sim}{A}_1{}^t\underset{\sim}{B}(T)] + \underset{\sim}{\delta}^t\underset{\sim}{h}\left\{\underset{\sim}{A}_1 + \underset{\sim}{M}^t[\chi(jT)]\underset{\sim}{A}_1\right\}\\ &- \tfrac{1}{2}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\,\underset{\sim}{M}^t[\chi(jT)]\underset{\sim}{A}_1\underset{\sim}{M}[\chi(jT)] - \underset{\sim}{\delta}^t\underset{\sim}{\mu}\,\underset{\sim}{M}^t[\chi(jT)]\underset{\sim}{A}_1\\ &\qquad\qquad - \underset{\sim}{M}^t[\chi(jT)]\underset{\sim}{P}\,\underset{\sim}{M}[\chi(jT)], \quad j=0,1,2\cdots.\end{aligned} \tag{47}$$

Since it is very tedious to determine whether $\underset{\sim}{R}[T,\chi(jT)]$ is positive definite on the basis of the general form used in obtaining (46), a particular coordinate system is now chosen.

If the original state variables are defined so that the output of the first plant is x_{11}, the output of the second plant is x_{21}, and in general the output of the i^{th} plant is x_{i1}, then

$$c(jT) = \sum_{i=1}^{N} x_{i1}(jT), \quad j=0,1,2,\cdots. \tag{48}$$

If, in addition,

---

*Because $\Delta\, V[\chi(jT)]$ for the linearized system is negative definite when $\underset{\sim}{Q}(T)$ is a positive definite matrix.

$$\underset{\sim}{A} \overset{\Delta}{=} \begin{array}{|c|c|c|c|} 1\ 0\ 0\ \cdots\ 0 & 1\ 0\ 0\ \cdots\ 0 & \cdots & 1\ 0\ 0\ \cdots\ 0 \\ 0\ 1\ 0\ \cdots\ \cdot & \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdots & \cdot\ \cdot\ \cdot\ \cdots\ 0 \\ 0\ 0\ 1\ \cdots\ \cdot & \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdots & \cdot\ \cdot\ \cdot\ \cdots\ 0 \\ \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdots & \cdot\ \cdot\ \cdot\ \cdots\ \cdot \\ \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdots & \cdot\ \cdot\ \cdot\ \cdots\ \cdot \\ \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdots & \cdot\ \cdot\ \cdot\ \cdots\ \cdot \\ 0\ 0\ 0\ \cdots\ \cdot & \cdot\ \cdot\ \cdot\ \cdots\ \cdot & \cdots & \cdot\ \cdot\ \cdot\ \cdots\ 1 \\ \hline n_1 \text{ columns} & n_2 \text{ columns} & & n_N \text{ columns} \end{array}, \qquad (49)$$

then

$$y_1(jT) = c\ (jT),\ j=0,1,2,\cdots, \qquad (50)$$

and

$$y_k(jT)=\begin{cases} x_{1k}(jT),\ k = 2,3,\cdots,n_1 \\ x_{2(k-n_1)}(jT),\ k = n_1+1,\ n_1+2,\cdots,n_1+n_2 \\ x_{3(k-n_1-n_2)}(jT),\ k=n_1+n_2+1,\ n_1+n_2+2,\cdots,n_1+n_2+n_3, \\ \qquad\qquad\qquad\qquad\qquad j=0,1,2,\cdots. \\ \qquad\qquad\cdot \\ \qquad\qquad\cdot \\ \qquad\qquad\cdot \\ x_{N(k-n+n_N)}(jT),\ k=n-n_N+1,\ n-n_N+2,\cdots,n \end{cases} \qquad (51)$$

Consequently, then,

$$\alpha_1 = 1, \qquad (52)$$

and

$$\alpha_i = 0,\ i = 2,3,\cdots n. \qquad (53)$$

It follows from (49), (52), and (53) that

$$
\underset{\sim}{R}[T,\underset{\sim}{\chi}(jT)] = \underset{\sim}{Q} +
\begin{bmatrix}
\begin{array}{l} 2\sum\limits_{i=1}^{n} b_{i1}\sum\limits_{j=1}^{n} p_{ij}m_{j1} \\ -\sum\limits_{i=1}^{n} m_{i1}\sum\limits_{j=1}^{n} p_{ij}m_{j1} \\ +\underset{\sim}{\delta}^{t}\underset{\sim}{h}(1+m_{11}) \\ -\underset{\sim}{\delta}^{t}\underset{\sim}{\mu}(m_{11}+\tfrac{1}{2}m_{11}^{2}) \\ +\underset{\sim}{\delta}^{t}\underset{\sim}{\mu}m_{11}b_{11}+\underset{\sim}{\delta}^{t}(\underset{\sim}{\mu}-\underset{\sim}{h})b_{11} \end{array}
&
\begin{array}{l} \sum\limits_{i=1}^{n} b_{i2}\sum\limits_{j=1}^{n} p_{ij}m_{j1} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}\underset{\sim}{\mu}m_{11}b_{12} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}(\underset{\sim}{\mu}-\underset{\sim}{h})b_{12} \end{array}
& \cdots\cdots &
\begin{array}{l} \sum\limits_{i=1}^{n} b_{in}\sum\limits_{j=1}^{n} p_{ij}m_{j1} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}\underset{\sim}{\mu}m_{11}b_{1n}+\tfrac{1}{2}\underset{\sim}{\delta}^{t}(\underset{\sim}{\mu}-\underset{\sim}{h})b_{1n} \end{array}
\\
\begin{array}{l} \sum\limits_{i=1}^{n} b_{i2}\sum\limits_{j=1}^{n} p_{ij}m_{j1} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}\underset{\sim}{\mu}m_{11}b_{12} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}(\underset{\sim}{\mu}-\underset{\sim}{h})b_{12} \end{array}
& 0 & 0\ \cdots & 0
\\
\cdot & \cdot & \cdots\cdot & \cdot \\
\cdot & \cdot & \cdots\cdot & \cdot \\
\cdot & \cdot & \cdots\cdot & \cdot \\
\begin{array}{l} \sum\limits_{i=1}^{n} b_{in}\sum\limits_{j=1}^{n} p_{ij}m_{j1} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}\underset{\sim}{\mu}m_{11}b_{1n} \\ +\tfrac{1}{2}\underset{\sim}{\delta}^{t}(\underset{\sim}{\mu}-\underset{\sim}{h})b_{1n} \end{array}
& 0 & 0\ \cdots & 0
\end{bmatrix},
\tag{54}
$$

where the arguments T and jT have been omitted to conserve space.

Now, for convenience, let

$$\underset{\sim}{Q} = \begin{bmatrix} q_{11} & 0 & 0 \cdots 0 \\ 0 & q_{22} & 0 \cdots 0 \\ \cdot & \cdot & \cdots\cdot 0 \\ \cdot & \cdot & \cdots\cdot 0 \\ \cdot & \cdot & \cdots\cdot 0 \\ 0 & 0 & 0 \cdots q_{nn} \end{bmatrix}, \tag{55}$$

where

$$q_{ii} > 0,\ i = 1,2,\cdots,n. \tag{56}$$

Then $\underset{\sim}{R}[T,\underset{\sim}{y}(jT)]$ is a positive definite matrix if and only if

$$\det\left\{\underset{\sim}{R}[T,\underset{\sim}{y}(jT)]\right\} > 0. \tag{57}$$

Evaluating the determinant gives

$$\det\left\{\underset{\sim}{R}[T,\underset{\sim}{y}(jT)]\right\} = \left[\prod_{i=2}^{n} q_{ii}\right]\Bigg\{q_{11}+2\sum_{i=1}^{n} b_{i1}\sum_{j=1}^{n} p_{ij}m_{j1}-\sum_{i=1}^{n} m_{i1}\sum_{j=1}^{n} p_{ij}m_{j1}$$
$$+\underset{\sim}{\delta}^t\underset{\sim}{h}(1+m_{11})-\underset{\sim}{\delta}^t\underset{\sim}{\mu}(m_{11}+\tfrac{1}{2}m_{11}^{\ 2})+\underset{\sim}{\delta}^t\underset{\sim}{\mu}\ m_{11}b_{11}+\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})b_{11}$$
$$-\sum_{k=2}^{n}\frac{1}{q_{kk}}\left[\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{n} p_{ij}m_{j1}+\tfrac{1}{2}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\ m_{11}b_{1k}+\tfrac{1}{2}\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})b_{1k}\right]^2\Bigg\},\ j=0,1,2,\cdots \tag{58}$$

$$= \left[a_0(\Delta K)^2 + a_1\ \Delta K + a_2\right]\prod_{i=2}^{n} q_{ii}, \tag{59}$$

where

$$a_0 \overset{\Delta}{=} -\left[\sum_{i=1}^{n} d_{i1}\sum_{j=1}^{n} p_{ij}d_{j1} + \tfrac{1}{2}d_{11}^{\ 2}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right]$$
$$-\left[\sum_{k=2}^{n}\frac{1}{q_{kk}}\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{n} p_{ij}d_{j1} + \tfrac{1}{2}\ b_{1k}d_{11}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right]^2 \tag{60}$$

and

$$a_1 \triangleq 2\sum_{i=1}^{n} b_{i1}\sum_{j=1}^{n} p_{ij}d_{j1} - 2K_0\sum_{i=1}^{n} d_{i1}\sum_{j=1}^{n} p_{ij}d_{j1}^{(1)}$$

$$+ d_{11}\underset{\sim}{\delta}^t(\underset{\sim}{h}-\underset{\sim}{\mu}) - K_0 d_{11}d_{11}^{(1)}\underset{\sim}{\delta}^t\underset{\sim}{\mu} + b_{11}d_{11}\underset{\sim}{\delta}^t\underset{\sim}{\mu}$$

$$- 2\sum_{k=2}^{n}\frac{1}{q_{kk}}\left[K_0\left(\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{n} p_{ij}d_{j1}^{(1)}+\tfrac{1}{2}d_{11}^{(1)}b_{1k}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right)\right.$$

$$\left.+\tfrac{1}{2}b_{1k}\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})\right]\left(\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{n} p_{ij}d_{j1}+\tfrac{1}{2}d_{11}b_{1k}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right) \qquad (61)$$

and

$$a_2 \triangleq q_{11}+\underset{\sim}{\delta}^t\underset{\sim}{h}+\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})b_{11}+K_0\left[d_{11}^{(1)}\underset{\sim}{\delta}^t\underset{\sim}{h}+d_{11}^{(1)}b_{11}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right.$$

$$\left.+ 2\sum_{i=1}^{n} b_{i1}\sum_{j=1}^{n} p_{ij}d_{j1}^{(1)} - d_{11}^{(1)}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right] - K_0^2\left[\sum_{i=1}^{n} d_{i1}^{(1)}\sum_{j=1}^{n} p_{ij}d_{j1}^{(1)}\right.$$

$$\left.+ \tfrac{1}{2}(d_{11}^{(1)})^2\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right] - \sum_{k=2}^{n}\frac{1}{q_{kk}}\left[K_0\left(\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{n} p_{ij}d_{j1}^{(1)}\right.\right.$$

$$\left.\left.+ \tfrac{1}{2}\, d_{11}^{(1)}b_{1k}\underset{\sim}{\delta}^t\underset{\sim}{\mu}\right) + \tfrac{1}{2}\, b_{1k}\underset{\sim}{\delta}^t(\underset{\sim}{\mu}-\underset{\sim}{h})\right]^2, \qquad (62)$$

with

$$\underset{\sim}{D}[\chi(jT)] \triangleq \begin{bmatrix} d_{11}[\chi(jT)] & d_{12}[\chi(jT)] & \cdots & d_{1n}[\chi(jT)] \\ d_{21}[\chi(jT)] & d_{22}[\chi(jT)] & \cdots & d_{2n}[\chi(jT)] \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ d_{n1}[\chi(jT)] & d_{n2}[\chi(jT)] & \cdots & d_{nn}[\chi(jT)] \end{bmatrix} = \begin{bmatrix} d_{11}[\chi(jT)] & 0 & 0\cdots 0 \\ d_{21}[\chi(jT)] & 0 & 0\cdots 0 \\ \cdot & & \cdot\ \cdots\cdots \\ \cdot & & \cdot\ \cdots\cdots \\ \cdot & & \cdot\ \cdots\cdots \\ d_{n1}[\chi(jT)] & 0 & 0\cdots 0 \end{bmatrix} \qquad (63)$$

and

$$\underset{\sim}{D}_1[\chi(jT)] \overset{\Delta}{=} \begin{bmatrix} d_{11}^{(1)}[\chi(jT)] & d_{12}^{(1)}[\chi(jT)] & \cdots & d_{1n}^{(1)}[\chi(jT)] \\ d_{21}^{(1)}[\chi(jT)] & d_{22}^{(1)}[\chi(jT)] & \cdots & d_{2n}^{(1)}[\chi(jT)] \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ d_{n1}^{(1)}[\chi(jT)] & d_{n2}^{(1)}[\chi(jT)] & \cdots & d_{nn}^{(1)}[\chi(jT)] \end{bmatrix} = \begin{bmatrix} d_{11}^{(1)}[\chi(jT)]00\cdots0 \\ d_{21}^{(1)}[\chi(jT)]00\cdots0 \\ \cdot \quad \cdots\cdots \\ \cdot \quad \cdots\cdots \\ \cdot \quad \cdots\cdots \\ d_{n1}^{(1)}[\chi(jT)]00\cdots0 \end{bmatrix}. \tag{64}$$

If the roots of the equation

$$a_0\, k^2 + a_1\, k + a_2 = 0 \tag{65}$$

are real and unequal, let the smaller root be denoted by k_1 and the larger, by k_2. Then, since $a_0 < 0$, $\det\left\{\underset{\sim}{R}[T,\chi(jT]\right\} > 0$ on and only on $K_0 + k_1 < K < K_0 + k_2$, $j = 0,1,2,\cdots$. If the roots of (65) are complex or real-and-equal, then there is no range of K on which $\det\left\{\underset{\sim}{R}[T,\chi(jT)]\right\} > 0$.

Since a_0, a_1, and a_2 are functions of $|y_1(jT)|$, k_1 and k_2 are functions of $|y_1(jT)|$. Let K_1 and K_2 denote the values of k_1 and k_2, respectively, obtained by maximizing k_1 with respect to $|y_1|$ and minimizing k_2 with respect to $|y_1|$. Then the nonlinear system is asymptotically stable in the large, for $r(t) \equiv 0$, at least on $(K_0 + K_1, K_0 + K_2)$. Upper and lower bounds on that range are K_a and K_b, where (K_a, K_b) is the range on which the linear system described by (46) is stable.

EXAMPLE

The use of the method presented in the earlier sections of this paper will now be demonstrated by applying the method to the pulse-width-modulated feedback system illustrated in Fig. 3, with

$$M = \alpha = T = 1. \tag{66}$$

In general, the output of the modulator is described by

$$m(t) = \begin{cases} m(jT), & 0 \le t-jT<\tau_j \\ 0, & \tau_j \le t-jT < T, \quad j=0,1,2\cdots, \end{cases} \tag{67}$$

where τ_j is the pulse width during the j^{th} sampling period. The modulator output is related to the modulator input by

$$m(jT) = M \operatorname{sgn}[r(jT) - c(jT)], \quad j=0,1,2,\cdots \tag{68}$$

and

$$\tau_j = T \operatorname{sat}\left[\frac{\alpha}{T}\,|r(jT) - c(jT)|\right], \quad j=0,1,2,\cdots, \tag{69}$$

where

$$\operatorname{sgn} x \triangleq \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \tag{70}$$

and

$$\operatorname{sat} x \triangleq \begin{cases} 1, & x \ge 1 \\ x, & 0 \le |x| < 1 \\ -1, & x \le -1 \end{cases} \tag{71}$$

The block diagram illustrated in Fig. 3 can be redrawn as shown in Fig. 4. It follows that, for the specified values of M, α, and T,

$$x_1[(j+1)] = e^{-1}x_1(j) + v_1(j), \quad j=0,1,2,\cdots \tag{72}$$

and

$$x_2[(j+1)] = e^{-2}x_2(j) - v_2(j), \quad j=0,1,2,\cdots, \tag{73}$$

where

$$\hat{v}_i(j) = -\frac{Ke^{-i}}{i}\left\{e^{i\,\operatorname{sat}[|x_1(j) + x_2(j)|]} - 1\right\} \operatorname{sgn}[x_1(j)+x_2(j)],$$

$$i=1,2; \; j=0,1,2,\cdots. \tag{74}$$

Use of (3) and (4) now gives

$$0 < h_i(j) = \frac{Ke^{-i}}{i} \left\{ \frac{e^{i\,\mathrm{sat}\left[\,|x_1(j) + x_2(j)|\,\right]} - 1}{|x_1(j) + x_2(j)|} \right\},$$

$$|x_1(j)+x_2(j)| \neq 0, \quad i=1,2; \; j=0,1,2,\cdots \tag{75}$$

and

$$\left| \frac{d\hat{v}_i(j)}{d[x_i(j)+x_2(j)]} \right| = \left| Ke^{-i} \left\{ e^{i\,\mathrm{sat}\left[\,|x_1(j)+x_2(j)|\,\right]} \right\} \right| \leq |K|,$$

$$i = 1,2; \; j=0,1,2,\cdots. \tag{76}$$

Letting

$$\underset{\sim}{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{77}$$

gives

$$\underset{\sim}{\Psi}(T) = \begin{bmatrix} e^{-1} & e^{-2} - e^{-1} \\ 0 & e^{-2} \end{bmatrix} \tag{78}$$

and

$$\underset{\sim}{D}[\underset{\sim}{y}(j)] = \begin{bmatrix} \dfrac{v_1[-y_1(j)] - v_2[-y_1(j)]}{|y_1(j)|} & 0 \\ \dfrac{v_2[-y_1(j)]}{|y_1(j)|} & 0 \end{bmatrix},$$

$$j=0,1,2\cdots. \tag{79}$$

To linearize the system, let

$$\underset{\sim}{D}_0 = \lim_{|y_1(j)| \to 0} \underset{\sim}{D}[\underset{\sim}{y}(j)] = K_o \begin{bmatrix} e^{-1} - e^{-2} & 0 \\ -e^{-2} & 0 \end{bmatrix} \tag{80}$$

The linearized system is then described by

$$\underset{\sim}{\chi}(j+1) = \begin{bmatrix} (1-K_0)e^{-1}+K_0e^{-2} & e^{-2}-e^{-1} \\ K_0e^{-2} & e^{-2} \end{bmatrix} \underset{\sim}{\chi}(j), \; j=0,1,2,\cdots. \tag{81}$$

It can readily be shown that such a system is stable on and only on

$$-2.34 < K_0 < 6.67. \tag{82}$$

To investigate the stability of the original nonlinear (pulse-width-modulated) system, let

$$K_0 = 6.00, \tag{83}$$

$$\underset{\sim}{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{84}$$

and

$$\underset{\sim}{\delta} = \begin{bmatrix} 0.4 \\ 0.4 \end{bmatrix}. \tag{85}$$

It is then found that

$$\underset{\sim}{P} = \begin{bmatrix} 10.08285 & 2.00753 \\ 2.00753 & 1.46739 \end{bmatrix}. \tag{86}$$

The values of $K_0 + k_1$ and $K_0 + k_2$ are plotted versus $|y_1|$ in Fig. 5. The maximum value of k_1 occurs for $|y_1| = 0.26$, and the minimum value of k_2 occurs for $|y_1| = 0.36$. The corresponding values of K_1 and K_2 are

$$K_1 = -6.281 \tag{87}$$

and

$$K_2 = -0.04. \tag{88}$$

Therefore, the pulse-width-modulated system is asymptotically stable in the large at least on

$$-0.281 < K < 5.96. \tag{89}$$

Since the linear approximation to the system is unstable for $K > 6.67$, the pulse-width-modulated system is not asymptotically stable in the large for $K > 6.67$.

CONCLUSION

The method presented in this paper can be used to obtain a sufficient condition for asymptotic stability in the large of nonlinear feedback systems of the kind illustrated in Fig. 1. It is not necessary that the poles of the transfer functions be real or simple, and the result obtained is not overly conservative, at least in the cases investigated to date.

REFERENCES

1. Kodama, S., "Stability of a Class of Discrete Control Systems Containing a Nonlinear Gain Element", Trans. IRE PGAC (Oct., 1962), pp. 102-109.
2. Pearson, J. B., and J. E. Gibson, "On the Asymptotic Stability of a Class of Saturating Sampled-Data System", Trans. IEEE, vol. 84, pt. II (Appl. and Ind.), pp. 81-86 (Feb., 1964).
3. Kalman, R. E., "Liapunov Functions for the Problem of Lur'e in Automatic Control", Proc. Nat'l. Acad. of Science, U.S.A., pp. 201-205 (Feb., 1963).
4. Popov, V. M., "Absolute Stability of Nonlinear Systems of Automatic Control", Automation and Remote Control, vol. 22, pp. 857-875 (Aug., 1961).

5. Tsypkin, Ya. Z., "Fundamentals of the Theory of Nonlinear Pulse Control Systems", Second Congress of IFAC (1964).

6. Jury, E. I., and E. B. Lee, "On the Stability of a Certain Class of Nonlinear Sampled-Data Systems", Trans. IEEE PGAC, vol. AC-9, pp. 51-61 (Jan., 1964).

7. Kalman, R. E., and G. P. Szego, "Sur la stabilite' absolue d'un systeme d'equations aux differences finies", Compt. rend. Acad. Sci, Paris, vol. 257, pp. 388-390 (July, 1963).

8. Szego, G. P., and J. B. Pearson, "On the Stability of Sampled-Data Systems: the Indirect Control Case", Trans. IEEE PGAC, vol. AC-9, no. 2, pp. 160-163 (April, 1964).

9. Ibrahim, E., and Z. V. Rekasius, "A Stability Criterion for Nonlinear Feedback Systems", Trans. IEEE PGAC, vol. AC-9, no. 2, pp. 154-159 (April, 1964).

10. Kalman, R. E., and J. E. Bertram, "Control System Analysis and Design via the Second Method of Liapunov, II", No. 59-NAC-3 Trans. ASME, Ser. D., vol. 82, pp. 349-400 (June, 1960).

ACKNOWLEDGEMENT

The work described in this paper was supported by the Air Force Office of Scientific Research, under Grant No. AF-AFOSR-270-63.

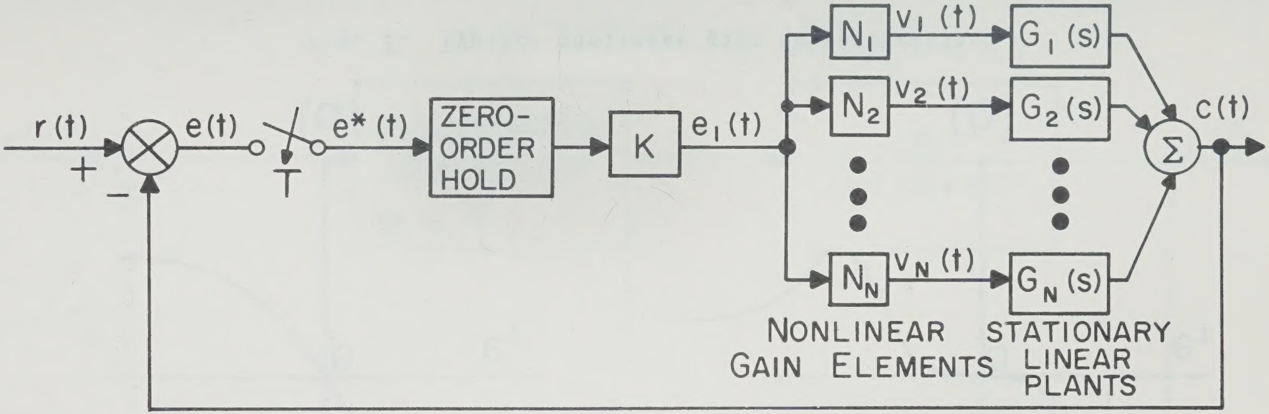

r(t)
+
-
e(t)
T
e*(t)
ZERO-ORDER HOLD
K
$e_1(t)$
N_1
N_2
N_N
$v_1(t)$
$v_2(t)$
$v_N(t)$
$G_1(s)$
$G_2(s)$
$G_N(s)$
Σ
c(t)
NONLINEAR GAIN ELEMENTS
STATIONARY LINEAR PLANTS

Fig. 1. A diagram of a sampled-data feedback system containing several nonlinear gain elements in parallel

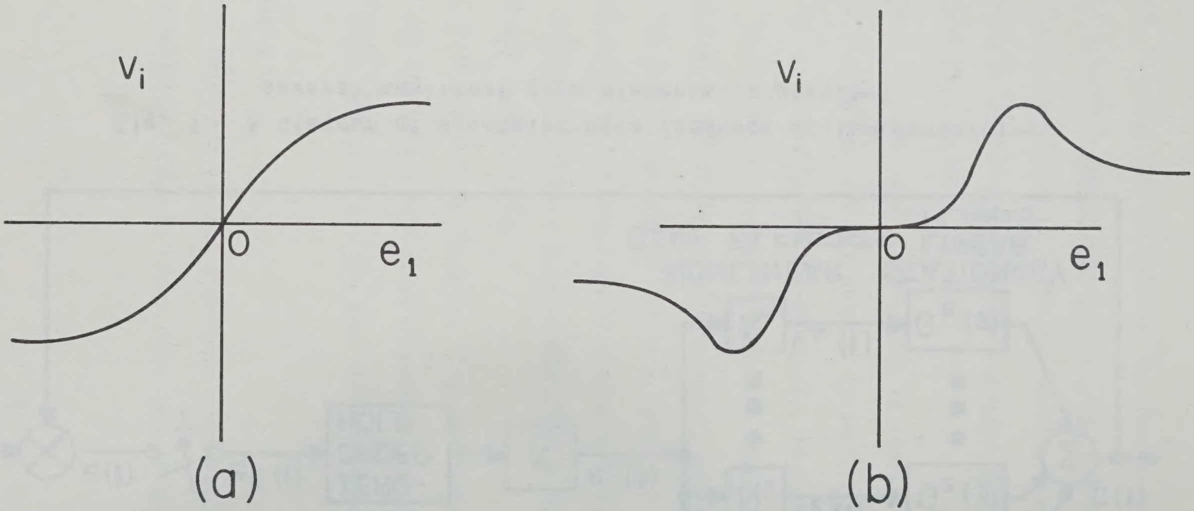


Fig. 2. Typical nonlinear gain characteristics

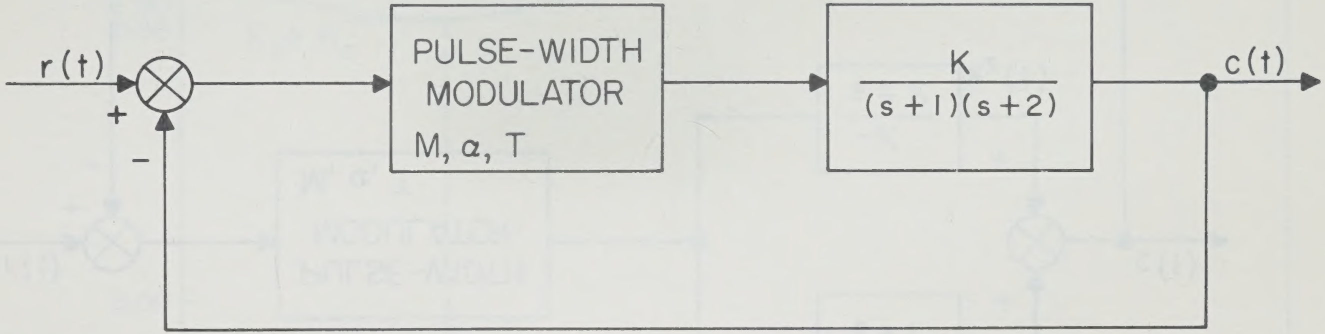

r(t)
+
−
PULSE-WIDTH
MODULATOR
M, α, T
$\frac{K}{(s+1)(s+2)}$
c(t)

Fig. 3. A block diagram of the pulse-width-modulated feedback system studied in the example

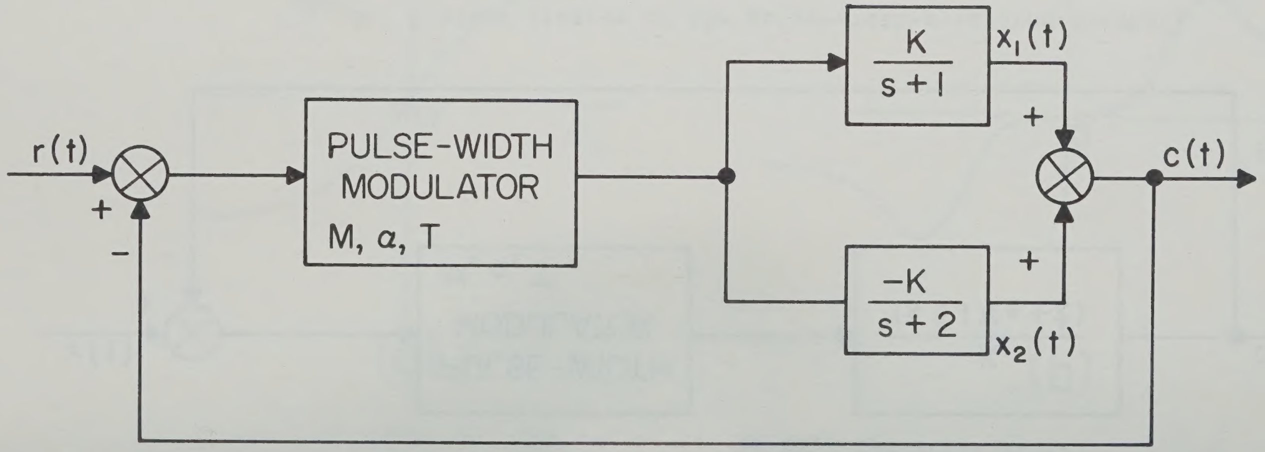

r(t)
+
−
PULSE-WIDTH
MODULATOR
M, α, T
$\frac{K}{s+1}$
$x_1(t)$
+
$\frac{-K}{s+2}$
$x_2(t)$
+
c(t)

Fig. 4. A block diagram equivalent to that shown in Fig. 3

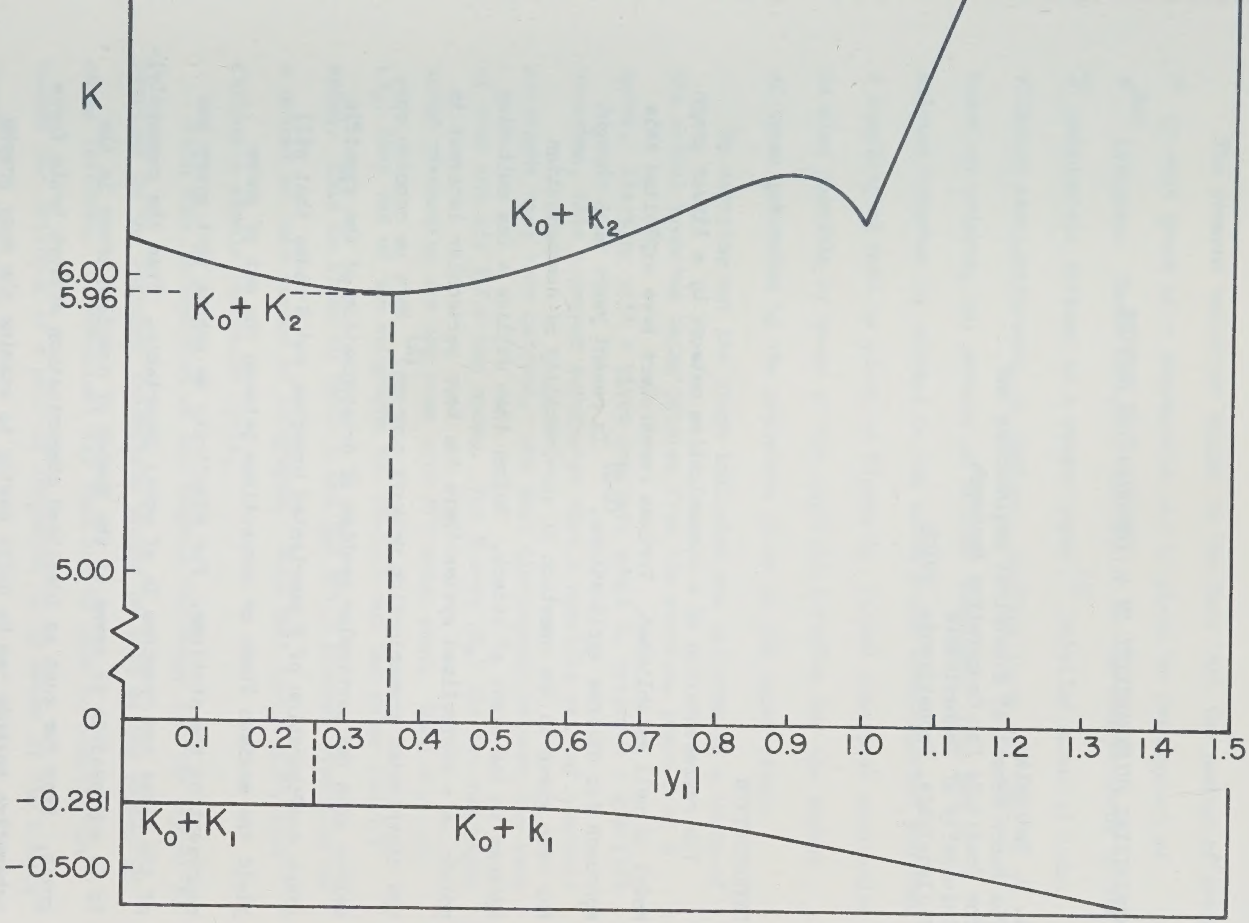


Fig. 5. A graph of critical gain versus magnitude of response

MINIMIZING VULNERABILITY IN A COMMUNICATION NETWORK

S. D. Bedrosian
The Moore School of Electrical Engineering and
The Institute for Cooperative Research
University of Pennsylvania
Philadelphia, Pennsylvania 19104

INTRODUCTION

The representation of a communication network by a linear graph model is well established. Various researchers have exploited this approach for diverse applications.(1-3) In recent years much thought has been given to the question of survivability of communication systems in the event of attack. Rather than utilize a few switching points in a centralized system there has been particular interest in the distributed communication network concept.(4) Here we concern ourselves with the particular problem of determination of the specific graph configuration of B unoriented branches and N nodes that will yield the maximum links or connections between the set of nodes representing the stations. For simplicity we assume that every one of the nodes and branches is of equal importance. Then the connectivity is expressible in terms of the number of complete trees in the graph. Even for such an idealized communication network brute force exhaustive methods can be quite costly to examine the many graphs specified only by N and B.

APPROACH

The present technique hinges on the fact that the number of trees T in each graph of a denumerable set is given by the product of N^{N-2} integers. These integers turn out to be the simple zeros of the $\mathfrak{F}$ polynomials defined in a recent paper.(5) Detailed proof of this, although straightforward, is tedious. Suffice it to say that it is based on ordering the pattern of deleted branches. This set of b deleted branches is related to the match stick combinatorial problem. A sampling of this is given in Figure 1. Actual numerical evaluation is also possible by means of the explicit formulas for the number of trees generated by the procedure given in the paper cited.

By carrying out the steps indicated one can compile a table of the actual branches being deleted from the complete graph G of N nodes. Starting with a given complete graph G, having $\widetilde{B} = N(N-1)/2$ branches, this approach terminates when a specific pair of complete subgraphs have been deleted. One must distinguish between the case of even and odd N for the nodes. For N even (N_e) the two subgraphs under discussion are the same with N/2 nodes each. When N is odd (N_o) then one of the subgraphs has $(N-1)/2$ and the other $(N+1)/2$ nodes. This is shown in Table 1 and for $N > 3$ focuses attention on a paired or twin series of subgraphs. Note that their appearance follows a simple recursion cycle.

With this information we can compute the range for the number of branches in the residual incomplete graphs as well as the number of cases involved. A convenient way that can be used to exhibit the bounds for deletion of a pair of subgraphs is by means of the Domain Chart(6) for connected linear graphs. Such a chart is a graphic representation of the fundamental relationship

Branches = Nullity + Rank.

It is shown in Figure 2 for graphs with up to 30 branches. The portion of the chart to the left of the heavy demarcation line indicating the upper bound includes the distinct connected linear (DCL) graphs, that is, connected graphs with no parallel branches and no branches connecting nodes to themselves. The remaining portion of the chart to the right of the heavy line pertains to graphs with at least one set of parallel branches. It is not implied that graphs with parallel branches cannot exist in the remaining portion of the chart. We are simply confining our attention to the class in the designated region. An immediate extension of the coverage just indicated may be obtained by application of the "image" principle discussed in a previous paper.(7) This permits us to map all the cases of incomplete graphs covered in Table 1 by pivoting each with respect to the complete graph G_N for the given N nodes to obtain their non-DCL counterparts. The range of graphs with $N > 3$ for this extended condition is shown in Figure 3.

DETAILED DISCUSSION

Note that the pattern repeats with centers at each of the higher order full graphs for $n > 1$. Therefore the number of cases of incomplete graphs which can be obtained from a complete graph equals the number of branches $\tilde{B}-b$ in the paired series of subgraphs. There is one case for each successive branch deleted up to

$N(N-2)/4$ cases for N_e

and

$(N^2-2N+1)/4$ cases for N_o.

The cases then fall in the range

$$\left.\begin{array}{l} N^2/4 \\ \\ (N^2-1)/4 \end{array}\right\} \leq B \leq N(N-1)/2 \qquad \begin{array}{l} \text{for } N_e \\ \\ \text{for } N_o \end{array}$$

where the maximum B is for the complete graph.

Simple explicit expressions can be given for the total number of trees for the extreme cases indicated in Figure 2.

For the complete graph

$$B = \tilde{B} = N(N-1)/2 \qquad \text{and } T_c = N^{N-2}$$

the lower extremes are:

$$\text{for } N_e \qquad B = \left(\frac{N}{2}\right)^2 \qquad \text{and } T = \left(\frac{N}{2}\right)^{N-2}$$

$$\text{for } N_o \qquad B = \frac{N^2-1}{4} \qquad \text{and } T = \left(\frac{N-1}{2}\right)^{\frac{N-1}{2}} \left(\frac{N+1}{2}\right)^{\frac{N-3}{2}}.$$

Working up from the lower extreme we can use the following more general formulas:

$$\text{for } N_e \quad \left(\frac{N}{2}\right)^2 \leq B \leq \frac{N}{2}\left(\frac{N}{2}+1\right)$$

$$T = \left(\frac{N}{2}+2\right)^{\rho}\left(\frac{N}{2}\right)^{N-2-\rho}$$

$$\rho = 0, 1, 2, \ldots \left(\frac{N}{2}-1\right).$$

Using the technique described in Reference 5 additional formulas can also be found for the intermediate cases. For graphs with $N \geq 9$ some additional remarks are needed to clarify just where the applicable upper limit of B lies for the "twin series." As in many card games we find that three of a kind supersedes a pair. This is shown for $N = 9$ in Figure 4. Observe that the limit of the twin series here is the four-cycle and five-cycle combination which has the formula

$$T = (N-4)(N-2)^2(N^2-5N+5)^2N^{N-9}.$$

The corresponding formula for deleting three three-cycles is

$$T = (N-3)^6N^{N-8}.$$

Upon evaluation of each of these for $N = 9$ we get 411,845 and 419,904 respectively. This confirms our previous statement. More than one such case will occur for $N > 9$.

CONCLUSIONS

A significant contribution has been made toward solution of the general problem of determining the graph with maximum connectivity, hence minimum vulnerability for the constraints specified. Having these explicit results we can contemplate less idealized(8) communication systems including those containing unreliable components(9-10) or those with lossy links.(11) Applications to other system problems also suggest themselves.

GLOSSARY OF TERMS

A "complete tree" is the minimum number (N-1) of branches which joins all the nodes N. Since we only use the complete tree in this paper the simpler term "tree" is used.

A "full graph" is defined as the interconnection of all pairs of nodes by exactly n branches. When n = 1 the full graph reverts to the well known "complete graph."

REFERENCES

1. C. Berge, The Theory of Graphs and Its Applications. London: Methuen and Co., English Translation, 1962.

2. S. Seshu and M. B. Reed, Linear Graphs and Electrical Networks. Reading, Mass: Addison-Wesley Publishing Co., Inc. 1961.

3. W. H. Kim and R. T. Chien, Topological Analysis and Synthesis of Communication Networks. New York: Columbia University Press, 1962.

4. P. Baran, "On Distributed Communications Networks," IEEE Trans. on Communications Systems, Vol. CS-12 (March 1964).

5. S. D. Bedrosian, "Generating Formulas for the Number of Trees in a Graph," Jour. Frank. Inst., Vol. 277, No. 4 (April 1964), pp. 313-326.

6. S. D. Bedrosian, "Application of Linear Graphs to Multilevel Maser Analysis," Jour. Frank. Inst., Vol. 274, No. 4, pp. 278-283 (Oct. 1962).

7. S. D. Bedrosian, "Properties of Linear Graphs Based on Trees," presented Fall URSI-IRE Meeting Ottawa, Canada, Oct. 1962.

8. K. Ikrath and H. Ulfers, "Topological Engineering of Communication Systems," IEEE Trans. Comm. Systems, Vol. CS-11, No. 1 (Mar. 1963), pp. 3-30.

9. I. M. Jacobs, "Connectivity in Probabilistic Graphs," MIT, RLE Tech. Report No. 356 (Sept. 15, 1959).

10. R. C. Amara, "Computer Design and Control of Probabilistic Communication Networks," Trans. On Communication Systems, Vol. CS-11 (Mar. 1963), pp. 30-35.

11. T. Fujisawa, "Maximal Flow in a Lossy Network," Proc. of Allerton Conf. on Circuit and System Theory, Urbana, Illinois (Nov. 1963), pp. 385-393.

ACKNOWLEDGEMENT

This paper is an outgrowth of work supported in part under contract number NObsr-91028 between BuShips U.S. Navy and I.C.R. University of Pennsylvania.

Fig. I Ordering of Deleted Subgraphs of b Branches

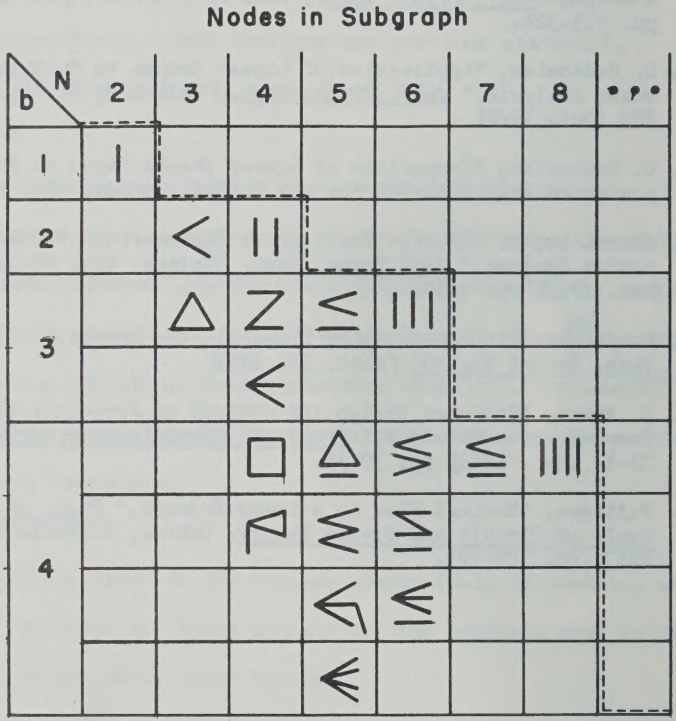
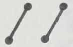
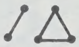
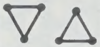
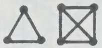
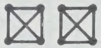
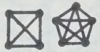
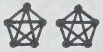

Nodes in Subgraph
N
b
2
3
4
5
6
7
8
•••
1
2
3
4

Table I Twin Series of Complete Subgraphs for Deletion from Graph Indicated

Nodes N		4	5	6	7	8	9	10	• • •	(even) N_e	(odd) N_o
Twin Series	Branches b deleted	2	4	6	9	12	16	20		$N(N-2)/4$	$(N^2-2N+1)/4$
	Sketch										
$\tilde{B}$-b	Branches in Residual Graph	4	6	9	12	16	20	25	• • •	$\frac{N^2}{4}$	$\frac{N^2-1}{4}$

$\tilde{B} = N(N-1)/2$

Fig. 2 Range of Graphs Covered by Use of the Twin Series of Table I

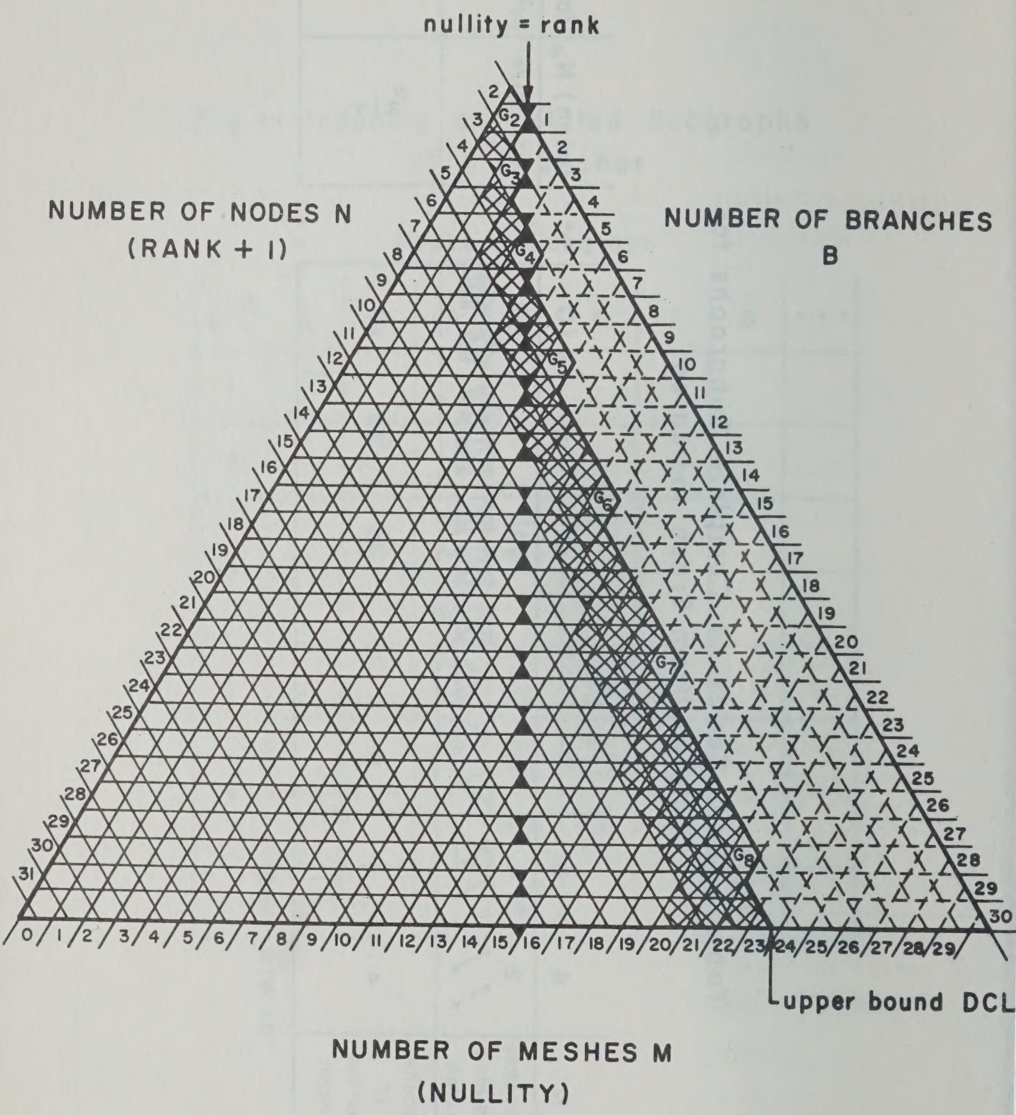


Fig. 3 Extended Range of Graphs for the Twin Series

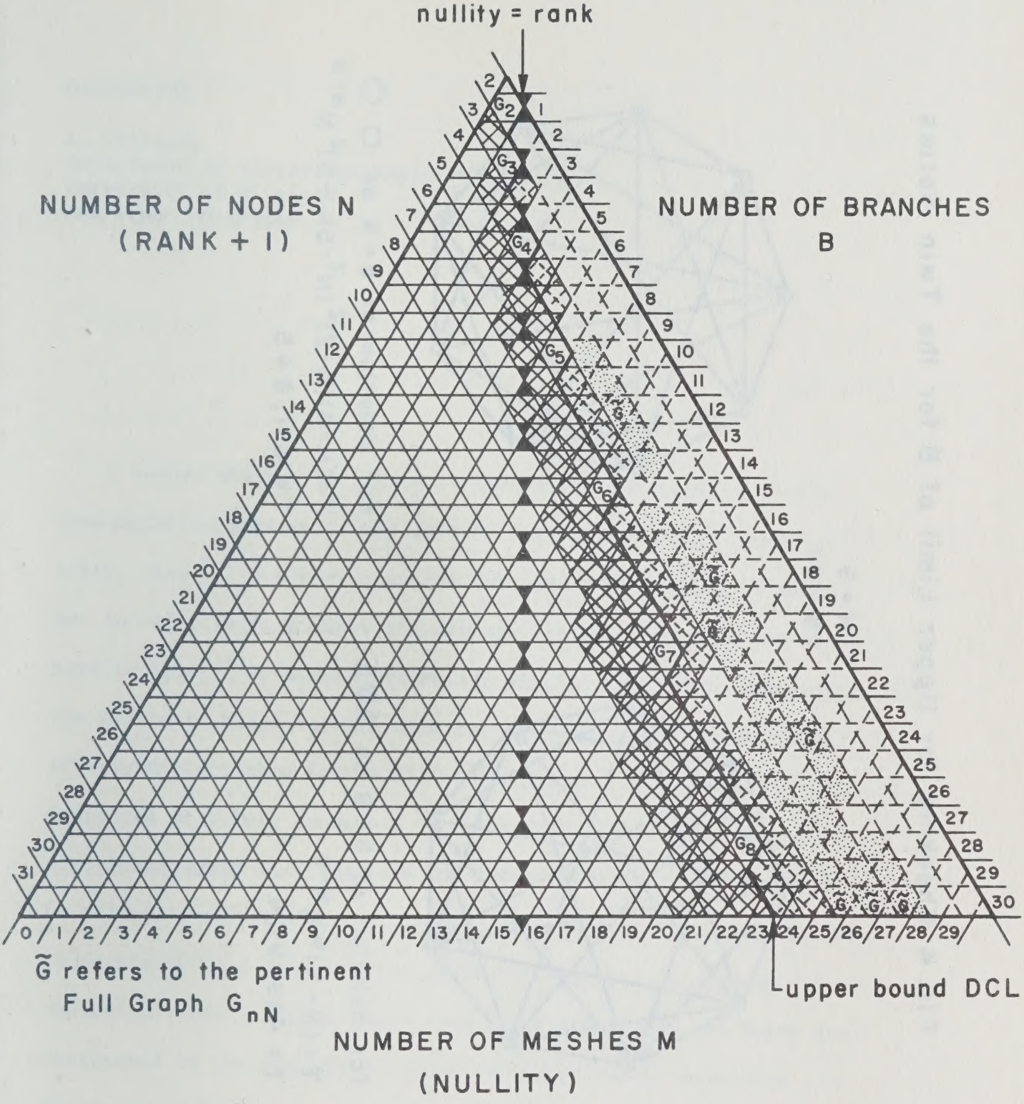
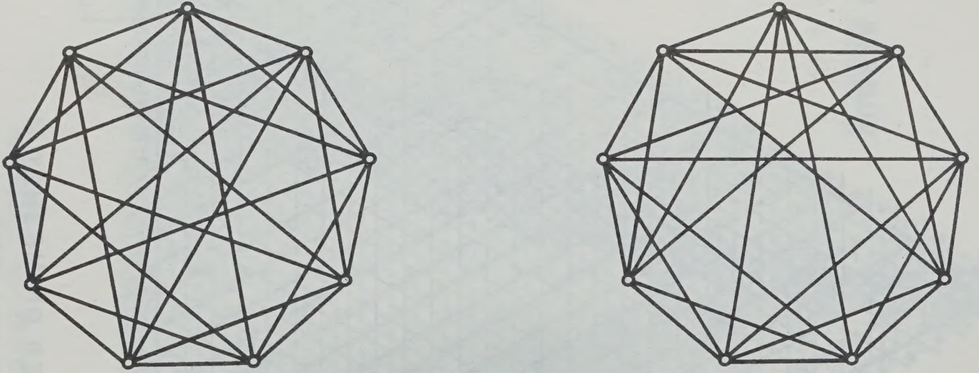


Fig. 4 Checking for Upper Limit of B for the Twin Series

$N = 9$
$B = 27$



for deletion of $b = 9$ as Δ∇Δ
$T = (N-3)^6 N^{N-8}$
$T = 419{,}904$

for deletion of $b = 9$ as □ ⬠
$T = (N-4)(N-2)^2(N^2-5N+5)^2 N^{N-9}$
$T = 411{,}845$

DISCUSSION

L. Weinberg
Department of Electrical Engineering
University of Michigan
Ann Arbor, Michigan

I wonder whether in trying to insure the survivability of a communication network we wouldn't be interested in its connectivity, that is, how many stations (or nodes) do we have to knock out in order to disconnect the network, or how many branches do we have to delete or block? There are simple theorems in terms of the distinct chains between each pair of nodes that should yield efficient algorithms for determining whether a graph is biconnected or triply connected or quintriply connected. These algorithms would also indicate the nodes or branches that should be deleted to disconnect the graph. For example, in a graph whose connectivity is 2 (that is, it is biconnected but not triply connected), two stations would have to be destroyed, and these are designated in the process of using the algorithm to determine the connectivity.

S. D. Bedrosian:

The observations made by Lou Weinberg certainly represent a point of view which may be adopted in treating the question of survivability. This is documented in Berge (Reference 1, Chapter 20).

On the other hand Ikrath (Reference 8) uses the more conventional definition for connectivity as the nullity or number of independent meshes and states that this is equivalent to the number of branches that can be removed from the graph without isolating a node from the others.

My paper was particularly addressed at the so-called distributed network approach exemplified by Baran (Reference 4). Baran used the term "redundancy level" as a measure of connectivity. This may be expressed as half the degree of the node k/2. Regarding node destruction the criterion of survivability he used is the percentage of stations not physically destroyed and remaining in communication with the largest single group of surviving stations. The combination of link and node destruction was indicated as being the worse case.

Perhaps it was not adequately emphasized that this was taken as the point of departure.

ON FLOWS IN COMMUNICATION NETWORKS

D. T. Tang
IBM Watson Research Center
Yorktown Heights, New York

INTRODUCTION

The concept of a communication network is a simple one. In addition to the topological configuration of the network, there is a real and positive number b_{ij} (capacity) associated with each edge (i, j). In a nonoriented communication network, b_{ij} and b_{ji} correspond to the same edge, and therefore $b_{ij} = b_{ji}$. They are in general distinct in oriented communication networks.

A flow $f_{ij}(x, y)$ from i to j can be thought of as the sum of path flow[(3)] components

$$f_{ij}(x, y) = \sum_{s} f_{ij}^{(s)}(x, y), \tag{1}$$

such that $f_{ij}^{(s)}(x, y) = v_{ij}^{(s)}$ if the edge (x, y) is on the path $P_{ij}^{(s)}$ from i to j, and $f_{ij}^{(s)}(x, y) = 0$ elsewhere. The value of the flow $f_{ij}(x, y)$ is

$$v_{ij} = \sum_{s} v_{ij}^{(s)} \quad . \tag{2}$$

When flows between different pairs of nodes exist simultaneously and are not to be confused with one another, they are referred to as multi-commody flows. A set of multi-commodity flows $F = \{ f_{ij}(x, y)\}$ is said to be feasible in a communication network B if

$$\sum_{f_{ij} \epsilon F} f_{ij}(x, y) \leqq b_{xy} \tag{3}$$

for all (x, y). In nonoriented cases, (3) becomes

$$\sum_{f_{ij} \epsilon F} [f_{ij} (x, y) + f_{ij} (y, x)] \leqq b_{xy} \tag{4}$$

A basic problem in the analysis of flows in networks is as follows: Given a set of flow requirements $\{ r_{ij} \}$ and a network B, what is the necessary and sufficient condition for a set of flows $\{ f_{ij}(x, y)\}$ with $v_{ij} = r_{ij}$ to be feasible in B?*

When a single commodity flow is considered, the above question is answered by the max-flow-min-cut theorem of Ford and Fulkerson(2) which applies to both oriented and nonoriented cases. When multi-commodity flows are considered, the situation becomes very involved and so far this challenging problem is only partially answered, as will be described in the following sections.

CUTS AND DISCONNECTING SETS

In order to generalize the max-flow-min-cut theorem to cover the feasibility of multi-commodity flows in various classes of networks, we need some more general notions than a minimum cut. First of all, a cut may be defined by

$$C (\alpha) = \{ (x, y) \mid x \epsilon \alpha, \ y \epsilon \bar{\alpha} \}, \tag{5}$$

---

*If flows $f_{ij}(x, y)$ with $v_{ij} = r_{ij}$ are feasible in B, the set of flow requirements is said to be feasible in or satisfied by B .

where α is a certain node set. With respect to each cut, there is a cut value:

$$V(\alpha) = \sum_{\substack{x \in \alpha \\ y \in \bar{\alpha}}} b_{xy} \,. \tag{6}$$

A disconnecting set in a network B, either oriented or non-oriented, is defined with respect to a number of k node pairs. Denoted by $C^s(i_1, \ldots i_k; j_1, \ldots, j_k)$, a disconnecting set is a collection of edges the removal of which from B destroys all paths from i_x to j_x, $x = 1, 2, \cdots, k$, with no proper subset of the collection having the same property. The value of a disconnecting set is the sum of all the corresponding edge capacities. A minimum disconnecting set is a disconnecting set assuming the minimum value among all existing ones with respect to the same set of node pairs. It is denoted by $C(i_1, \ldots, i_k; j_1, \ldots, j_k)$ and its value by $V(i_1, \ldots, i_k; j_1, \cdots, j_k)$. When $k = 1$, a minimum disconnecting set becomes a minimum cut.

The value of a minimum disconnecting set with respect to a given set of node pairs can be obtained by means of a systematic procedure based on the following theorem.

Theorem 1:

$$V(i_1, \ldots, i_k; j_1, \ldots, j_k) = \min \left[\text{values of } \bigcup_{\substack{x \\ s}} C^s(i_x; j_x) \right] \tag{7}$$

where the minimum is taken over all possible combinations of different cuts separating node pairs (i_x, j_x) for $x = 1, 2, \ldots, k$.

Proof:

For any collection of k cuts of B, $C^s(i_x; j_x)$, $x = 1, \ldots, k$, the removal of the union $\bigcup_{x=1}^{k} C^s(i_x; j_x)$ clearly destroys all paths directed from i_x to j_x. Therefore,

$$\bigcup_{x=1}^{k} C^{s}(i_x;\, j_x) \supseteq C^{s}(i_1, \dots, i_k;\, j_1, \dots, j_k). \qquad (8)$$

On the other hand, given any disconnecting set, $C^{t}(i_1, \dots, i_k;\, j_1, \dots, j_k)$, its removal destroys all paths directed from i_x to j_x. It follows that

$C^{t}(i_1, i_2, \dots, i_k;\, j_1, j_2, \dots, j_k)$ contains some $C^{t}(i_x;\, j_x)$ for every x, $x = 1, 2, \dots, k$, and therefore contains the union of some collection of cuts $\bigcup_{x=1}^{k} C^{t}(i_x;\, j_x)$. From the first part of the proof, $\bigcup_{x=1}^{k} C^{t}(i_x;\, j_x)$ contains some $C^{s}(i_1, \dots, i_k;\, j_1, \dots, j_k)$. That is,

$$C^{t}(i_1, \dots, i_k;\, j_1, \dots, j_k) \supseteq \bigcup_{x=1}^{k} C^{t}(i_x;\, j_x) \supseteq C^{s}(i_1, \dots, i_k;\, j_1, \dots, j_k) \quad ($$

Since $C^{t}(i_1, \dots, i_k;\, j_1, \dots, j_k)$ is minimal by definition, we have

$$C^{t}(i_1, \dots, i_k;\, j_1, \dots, j_k) = \bigcup_{x=1}^{k} C^{t}(i_x;\, j_x) \qquad (10)$$

for each disconnecting set. From (8) and (10), the minimum disconnecting set value is

$$V(i_1, \dots, i_k;\, j_1, \dots, j_k) = \min \left[\text{values of } \bigcup_{\substack{x \\ s}} C^{s}(i_x;\, j_x) \right], \qquad (11)$$

which proves the theorem.

From the max-flow-min-cut theorem, a single flow requirement is feasible in a network B if, and only if, $V(i;\, j) \geqq r_{ij}$. (12)

When a set of multi-commodity flow requirements $\{r_{ij}\}$ is given, two sets of inequalities can be written as the generalized conditions of (12).

The first condition is in terms of cut values; it reads:

$$V^{b}(\alpha) \geqq V^{r}(\alpha) \qquad (13)$$

for all cuts of B, where $V^{b}(\alpha)$ is evaluated in B and $V^{r}(\alpha)$ is evaluated in R , the network with r_{ij} as edge capacities.

The second condition is in terms of disconnecting set values; it reads:

$$V^b(i_1, \ldots, i_k; j\,, \ldots, j_k) \geqq \begin{matrix} V^r(i_1, \ldots, i_k; j_1, \ldots, j_k) \\ \sum_{x=1}^{k} r_{i_x j_x} \end{matrix} \qquad (14)$$

for all disconnecting sets with respect to all distinct collections of node pairs (i_x, j_x) in R .

The relationship between (13) and (14) in both nonoriented and oriented cases will next be discussed.

NONORIENTED CASES

When nonoriented networks and flow requirements are considered, we have the following theorems:

Theorem 2:

A given set of two flow requirements is feasible in a nonoriented network B if, and only if, condition (14) holds.

Theorem 2 was first obtained by Hakimi.(5) It has been found that the theorem does not apply to situations with three (or more)-commodity flows.(6)

Theorem 3:

In nonoriented cases, conditions (13) and (14) are equivalent.

Theorem 3 was proved by Tang.* It can be shown that, comparing to (14), (13) is usually a much more compact set of inequalities requiring far less computation.

Recently, a special class of networks called "bi-path" networks was studied.(11) Such a network is characterized by its property that each of its maximal nonseparable components with five or more nodes contains a Hamiltonian circuit with respect to which there exist no cross-over connections.** From the topological properties of bi-path networks, the following theorem has been established.

---

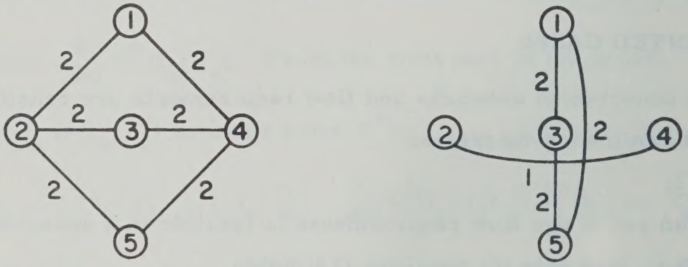
* See the proof of Theorem 2 in Reference 9. The statement of Theorem 2 in Reference 9 should be replaced by the statement of Theorem 3 in this paper.

Theorem 4:

A given set of multi-commodity flow requirements is feasible in a bi-path network if, and only if, condition (13) holds.

In the following example, a set of flow requirements (Figure 1-B) is not feasible in the network shown in Figure 1-A although condition (13) is satisfied.(6) This network is the simplest network which is not bi-path.

Example 1:



(A) B-network (B) R-network

Figure 1

The class of bi-path networks is the largest class of nonoriented networks in which Theorem 4 applies. For each network which is not bi-path, there exists a set of flow requirements which is not feasible in the network, although condition (13) is satisfied. Such counter-examples can easily be constructed from Example 1 by noting that each network which is not bi-path contains a subnetwork with the configuration shown in Figure 1-A.

ORIENTED CASES

The feasibility conditions of multi-commodity flows is perhaps made even more difficult in oriented networks. The cut value

---

** A cross-over consists of two edges (i, j) and (a, b) such that the terminal nodes of these two edges alternate in their order of appearance around the circuit.

condition of (13) and the disconnecting set condition of (14) are no longer equivalent in general. Condition (14) now becomes a stronger set of inequalities implying (13). The following theorems establish classes of oriented networks in which either (13) or (14) remains as necessary and sufficient feasibility condition for multi-commodity flows:

Theorem 5:

A set of multi-commodity flow requirements is feasible in a full duplex tree* if, and only if, condition (13) is satisfied for 2 (n-1) cuts each of which corresponds to an oriented edge of the full duplex tree.

Proof:

The necessity of condition (13) is immediate and requires no proof here. The sufficiency can be seen from the following constructive procedure: For any given $r_{ij} > o$, there exists a unique path P_{ij} from i to j in the full duplex tree. A feasible $f_{ij}(x, y)$ must therefore contain the only path flow component along P_{ij} with value $v_{ij} = r_{ij}$. It is clear that since all these paths are unique, such flows are feasible if condition (13) holds for 2 (n-1) cuts corresponding to the oriented edges of the duplex tree.

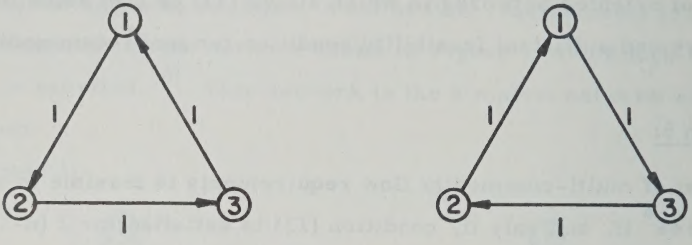
The following example shows a set of flow requirements (Figure 2-B) which is not feasible in the network shown in Figure 2-A, although condition (13) is satisfied. This network is the simplest strongly connected oriented network** which is not a full duplex tree.

---

* A full duplex tree is an oriented network of n nodes and 2(n-1) edges in which edges (i, j) and (j, i) always appear in pairs. When replacing all the oriented edge pairs by single nonoriented edges, the resultant network becomes a tree.

** A strongly connected network is one which contains a path from i to j for every distinct ordered node pair.

Example 2:



(A) B-network (B) R-network

Figure 2

Among strongly connected oriented networks, the class of full duplex trees is the largest class of networks in which Theorem 5 applies. It can easily be seen that a strongly connected oriented network either is a full duplex tree, or contains an oriented loop[*] of three or more edges. Counter-examples can than easily be constructed from Example 2.

It is interesting to note that the proof of Theorem 5 depends only upon the facts that (i) the path between each ordered node pair is unique, and (ii) each edge constitutes a cut which gives rise to an inequality in condition (13).

Thus by considering path-unique oriented networks in general, we still have the desired property that each edge alone constitutes a disconnecting set which gives rise to an inequality in condition (14).

A path-unique oriented network is characterized by the following theorem.

Theorem 6:

An oriented network B which contains a unique path between each ordered node pair consists of connected oriented loops such that any two loops can have at most one node in common, and if each loop is replaced by a nonoriented subtree connecting the nodes on the loop, the resulting network B' is a tree.

---

[*] A loop in an oriented network is a closed oriented path.

Proof:

Let (i, j) be any edge in B, than B must also contain a path from j to i. Therefore (i, j) belongs to some loop in B . It follows that every edge in B belongs to some loop in B . Two different loops in B cannot have two nodes in common, for assuming a and b to be two such nodes, then either two different paths exist from a to b or b to a; or two loops must completely coincide. In either case contradiction results. If we replace each loop by a nonoriented subtree, the resulting network B' must be connected, for otherwise there exists a certain node pair (x, y), such that there is no path from x to y in B. Furthermore, B' does not contain any circuit, for if it does, there exist a large loop in B corresponding to the circuit in B', which has at least two nodes in common with each loop corresponding to a subtree, a situation ruled out earlier in this proof. Hence B' must be a tree.

From an earlier discussion, we now have the following theorem:

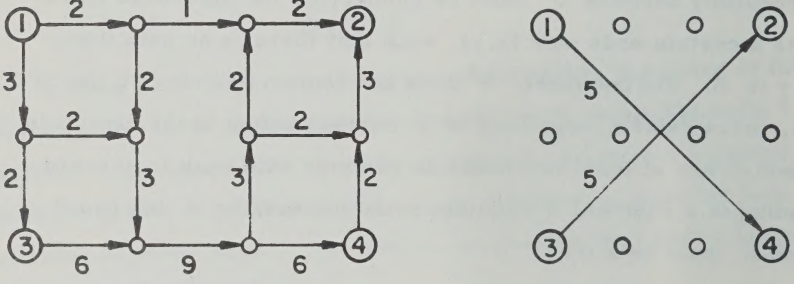
Theorem 7:

A set of multi-commodity flow requirements is feasible in a path-unique oriented network if, and only if, condition (14) is satisfied for all disconnecting sets each of which corresponds to an edge in the network.

In view of the limited classes of oriented networks obtained above in which the necessary and sufficient condition for the feasibility of arbitrary multi-commodity flows takes the form of either (13) or (14), a natural question to ask at this point is whether (13) or (14) would indeed be necessary and sufficient for the feasibility of k-commodity flows in an arbitrary oriented network if k is a small integer.

Surprisingly, the answer to the above question is that either (13) or (14) is necessary and sufficient for k-commodity flows to be feasible in an arbitrary oriented network only when k = 1, which is the case equivalent to the max-flow-min-cut theorem. The following example shows a set of two flow requirements which is not feasible in the network shown in Figure 3-A, although condition (14), and hence (13), is satisfied.

Example 3:



(A) B-network

(B) R-network

Figure 3

TIME-SHARING VS CAPACITY-SHARING

When a communication network is to satisfy simultaneous flow requirements, multi-commodity flows between different node pairs are usually expected to exist simultaneously. Such a scheme allows sharing of channel capacity of any edge in the network for multi-commodity flows. A quite different approach based on single-commodity concepts leads to the time-sharing scheme which allows only a single flow between a certain node pair at any specific instant. Different requirements are satisfied by proper sharing in time.

Consider, in either oriented or nonoriented cases, that a network B with terminal matrix[8] $\bar{T} = [t_{ij}]$ is operated on time-sharing basis, such that in each unit time period, only w_{ij} is used for the flow from i to j. The only constraint on w_{ij} is

$$\sum_{(i,j)\,\epsilon R} w_{ij} = 1 , \tag{15}$$

where R is the requirement network constructed from the given set of flow requirements. In order for the requirements to be satisfied by B , we must have

$$r_{ij} \leqq w_{ij}\, t_{ij} \tag{16}$$

for all $(i, j)\ \epsilon R$.

We now prove the following theorem:

Theorem 8:

If a given set of flow requirements represented by R is satisfied by a communication network B by time-sharing, then R is feasible in B in the form of multi-commodity flows by capacity-sharing. This applies to both oriented and nonoriented cases.

Proof:

Let $\bar{T} = [t_{ij}]$ be the terminal matrix of B , and let R represent the given set of flow requirements which is satisfied by B by time-sharing, then there exists w_{ij}'s satisfying (15) and (16). Consider networks $B_{ij} = w_{ij}\, B^{*}$ for all w_{ij} as component networks of B , then

$$\sum_{(i,j)\,\epsilon R} B_{ij} = \sum_{(i,j)\,\epsilon R} w_{ij} B = B \sum_{(i,j)\,\epsilon R} w_{ij} = B . \tag{17}$$

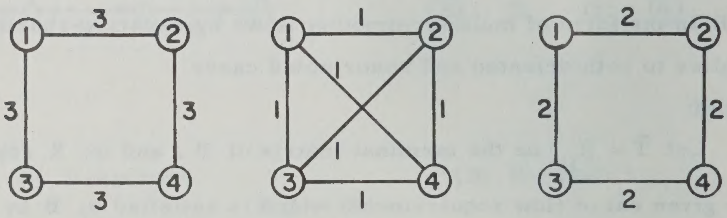
---

* i.e., B_{ij} consists of edges (x, y) with capacities $w_{ij}\, b_{xy}$.

The terminal matrix for B_{ij} is $w_{ij} \bar{T}$ with the (i, j)th entry equal to $w_{ij} t_{ij}$. But $w_{ij} t_{ij}$ is no smaller than r_{ij} from (16). Since this is true for all node pairs, multi-commodity flows for all r_{ij} exist in $B = \Sigma B_{ij}$ by capacity sharing.

Theorem 8 proves in general the superiority of the capacity-sharing scheme over the time-sharing scheme in network economy or operating efficiency when simultaneous flow requirements are to be satisfied by the network. However, the multi-commodity flows formulated in the above proof is usually not very economical. This point is illustrated in the following example with nonoriented network and flow requirements.

Example 4:



(A) B-network (B) R-network (C) B'-network

Figure 4

The network in Figure 4-A has the terminal matrix

$$\bar{T} = \begin{bmatrix} \textcircled{1} & 6 & 6 & 6 \\ 6 & \textcircled{2} & 6 & 6 \\ 6 & 6 & \textcircled{3} & 6 \\ 6 & 6 & 6 & \textcircled{4} \end{bmatrix} \tag{18}$$

With $w_{ij} = \frac{1}{6}$ for all $(i, j) \epsilon R$, (15) and (16) are both satisfied. Since $B_{ij} = \frac{1}{6} B$ and $r_{ij} = 1$, the flow $f_{ij}(x, y)$ clearly saturates all the edge capacities in B_{ij}. Therefore B is completely saturated by the multi-commodity flows established in this manner.

A different set of multi-commodity flows exists which only uses a total of 2 out of 3 units of each edge capacity in B (Figure 4-A). This may be verified by applying Theorem 4 to B' (Figure 4-C)and R (Figure 4-B), noting that B and B' are bi-path.

When the total branch capacity is to be minimized in a non-oriented case, the best time-sharing scheme possible to satisfy a given set of flow requirements R is to have a realizable terminal matrix $\bar{T}$ and a set of w_{ij} satisfying (15) and (16) with equality signs. The optimal realization of $\bar{T}$ can be obtained by the procedure of Wing and Chien, (13) which is equivalent to the procedure of Gomory and Hu(4) written in partitioned form.(8) This optimal realization can be shown to be bi-path in configuration. An efficient set of multi-commodity flows can thereby be established.

The under-saturated edge capacities thus give a measure to the saving of total capacities using the capacity-sharing scheme with multi-commodity flows.

CONCLUSIONS

We have considered in this paper some special cases and related topics of the challenging problem — to find a general necessary and sufficient conditions for the multi-commodity flows to be feasible in a communication network. Classes of oriented and nonoriented networks in which the set of inequalities on cut calues and the set of inequalities on minimum disconnecting set values become necessary and sufficient feasibility conditions are presented. An important example which demonstrates the insufficiency of minimum disconnecting set conditions for the feasibility of two-commodity flows in an oriented network is included. We have also proved that, in satisfying a given set of flow requirements, the single-commodity time-sharing scheme is never better than the multi-commodity capacity-sharing scheme from the viewpoint of network economy.

It is hoped that results presented in this paper will help in the understanding and the general solution of this complex problem.

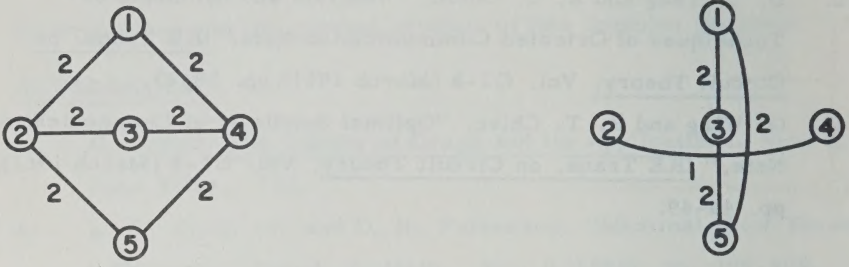
REFERENCES

1. C. Berge, The Theory of Graph and its Applications. New York: John Wiley, 1962.
2. L. R. Ford, Jr. and D. R. Fulkerson, "Maximal Flow Through a Network," Canad. J. Math., Vol. 8 (1956), pp. 399-404.
3. L. R. Ford, Jr. and D. R. Fulkerson, Flows in Networks. Princeton, N. J.: Princeton Univ., 1962.
4. R. E. Gomory, and T. C. Hu, "Multi-Terminal Network Flows," J. SIAM., Vol. 5 (December 1961), pp. 551-570.
5. S. L. Hakimi, "Simultaneous Flows Through a Communication Network," IRE Trans. on Circuit Theory, Vol. CT-9 (June 1962), pp. 169-175.
6. T. C. Hu, "Multi-Commodity Network Flows," J. of Operations Research, Vol. 11 (May 1963), pp. 344-360.
7. W. H. Kim, and R. T. Chien, Topological Analysis and Synthesis of Communication Networks, New York: Columbia Univ. Press, 1962.
8. W. Mayeda, "Terminal and Branch Capacity Matrices of a Communication Net," IRE Trans. on Circuit Theory, Vol. CT-7 (September 1960), pp. 261-269.
9. D. T. Tang, "Communication Networks with Simultaneous Flow Requirements," IRE Trans. on Circuit Theory, Vol. CT-9 (June 1962), pp. 176-182.
10. D. T. Tang, "Optimal Trees for Simultaneous Flow Requirements," Proceedings of the NEC, Vol. 19 (October 1963), pp. 28-32.

11. D. T. Tang, "Bi-Path Networks and Multi-Commodity Flows," to appear in IEEE Trans. on Circuit Theory, December 1964.

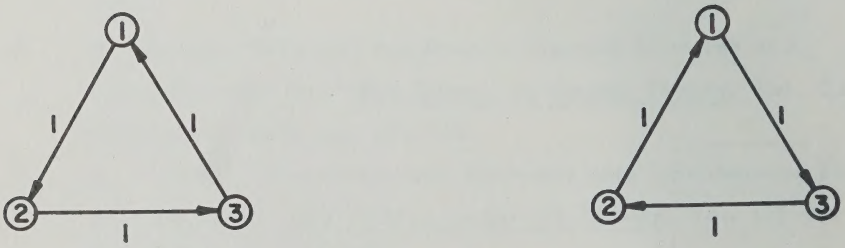
12. D. T. Tang and R. T. Chien, "Analysis and Synthesis of Techniques of Oriented Communication Nets," IRE Trans. on Circuit Theory, Vol. CT-8 (March 1961), pp. 39-43.

13. O. Wing and R. T. Chien, "Optimal Synthesis of Communication Nets," IRE Trans. on Circuit Theory, Vol. CT-8 (March 1961), pp. 44-49.


1
2
2
2
3
4
2
2
5
2
2
1
2
3
2
4
1
2
5

(A) (B)

Fig. 1


1
1
1
2
3
1
1
1
1
2
3
1

(A) (B)

Fig. 2

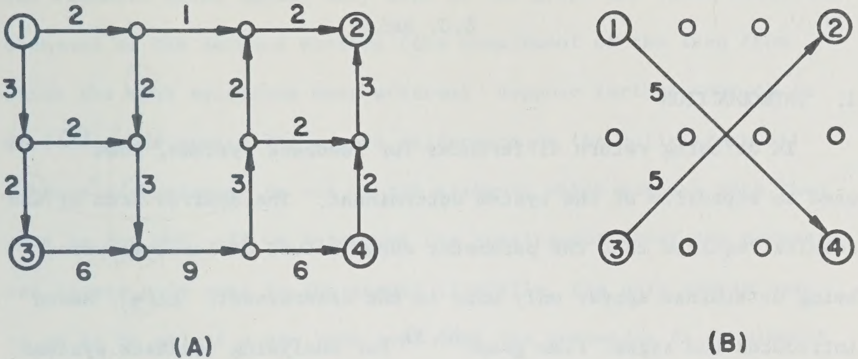

(A)
(B)

Fig. 3

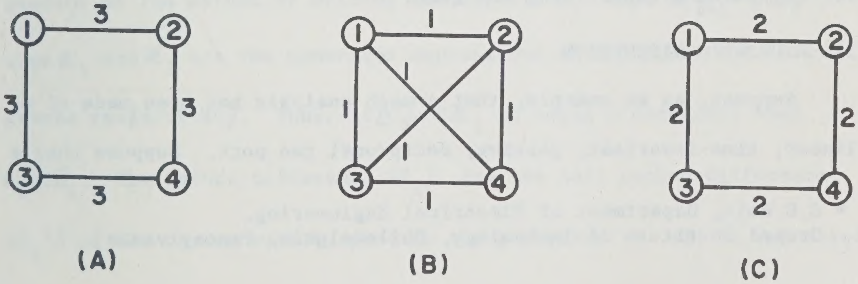

(A)
(B)
(C)

Fig. 4

MULTI-RETURN DIFFERENCES OF FEEDBACK SYSTEMS

G.C. Reis*

I. INTRODUCTION

In defining return differences for feedback systems, Bode(1) used an expansion of the system determinant. The desired form of his results required that the parameter whose return differences were being determined appear only once in the determinent. Later, Mason introduced the signal flow graph(2,3) for analyzing feedback systems. The signal flow graph (SFG) rapidly became the most popular means for feedback analysis. Bode's original requirement that the parameter of interest appear only once in the system determinant was translated into SFG language as requiring that the parameter appear only once in the system graph. This has been stated many times by many authors,(4,5,6,7) including the present one.(8,9,10,11) Taken at face value, this requirement is too stringent. It is the purpose of this paper to show when this requirement can be relaxed. Since sensitivity (in the Bode sense) and system gain (by Blackman's equation) are expressed in terms of return differences, the additional freedom provided by the results of this paper yield simple methods for determining sensitivity and gain.

II HEURISTIC DISCUSSION

Suppose, as an example, that a mesh analysis has been made of a linear, time-invariant, passive, reciprocal two port. Suppose that a

---

* G.C.Reis, Department of Electrical Engineering,
Drexel Institute of Technology, Philadelphia, Pennsylvania

signal flow graph has been constructed in the standard fashion.(2,3) The elements which appear only once in the graph and those which are elements of the network co-tree (the complement of the tree from which the mesh equations were written) Suppose further that it is desired to determine the return difference or the null-return difference with respect to one of the elements which appears more than once in the SFG. If we interpret the requirement (that the parameter can appear only once in the graph) literally, the only course open to us is to select a new tree such that the parameter is an element of the co-tree, and construct a new graph. Suppose that this has been done. Let the determinant of the first graph be Δ_1, and that of the second, Δ_2. Let interest be centered on the voltage transfer ratio (E_2/E_1) of the two port. By use of Mason's Gain Formula $\Sigma_2/\Sigma_1 = \frac{\Sigma}{\Delta}$, where Δ is the system determinant and Σ is the sum of products of forward transmittances by the determinant left when each is removed, as described by Mason.(2) The following observations can then be made.

The stability of the system is determined from the roots of the determinant Δ. Since this is independent of the way the system equations are written, Δ_1, can differ from Δ_2 by, at most, a multiplicitive constant. Similarly, the voltage transfer ratio is independent of the method of writing the equations. Thus $\Sigma_1/\Delta_1 = \Sigma_2/\Delta_2$, when Σ_1 and Σ_2 are the numerator expressions of the first and second graphs respectively. Thus, if $\Delta_2 = C\Delta_1$ (C being a constant) then $\Sigma_2 = C\Sigma_1$. The return difference (F_k) and the null return difference (F_k') of the system with respect to the parameter k can be expressed in general as$^{(10)}$

$$F_k = \frac{\Delta}{\Delta^o} \tag{1}$$

$$F_k' = \frac{\mathcal{E}}{\mathcal{E}^o} \tag{2}$$

where the superscript o indicates that k is set equal to zero. Let F_{2k} and F_{2k}' be the return difference, and null return differencee, respectively as determined from the second SFG. Then

$$F_{2k} = \frac{\Delta_2}{\Delta_2^o} = \frac{c\Delta_1}{c\Delta_1^o} = \frac{\Delta_1}{\Delta_1^o} = F_{1k} \tag{3}$$

$$F_{2k}' = \frac{\mathcal{E}_2}{\mathcal{E}_2^o} = \frac{c\mathcal{E}_1}{c\mathcal{E}_1^o} = \frac{\mathcal{E}_1}{\mathcal{E}_1^o} = F_{1k}' \tag{4}$$

where F_{1k} and F_{1k}' indicate the return difference and null return difference, respectively, as determined from the first graph. However, the parameter k was assumed to appear more than once in the first graph. Hence, one is led to believe that the application of definitions (1) and (2) will yield the return differences, whether the parameter appears once, or more than once in the SFG.

III PROOF FOR SINGLE PARAMETER CASE

In this section it will be shown that the results of the preceding example are rather general, requiring only that the system determinant be linear in the parameter k and that $\mathcal{E}^o \neq 0$ Since the utility of these results will most likely be in the calculation of gain and sensitivity, it will also be shown that these considerations also apply to F_k and F_k' defined as in equations (5) and (6)$^{(8)}$

$$F_k = \frac{\mathcal{E}_I}{\mathcal{E}_I^o} \tag{5}$$

$$F_k' = \frac{\Delta_I}{\Delta_I^o} \tag{6}$$

The subscript I refers to the graph obtained by inverting one forward path of the given graph. The superscript zero again indicates that k has been set to zero. The gain (G_I) of this inverted graph is Σ_I/Δ_I. By definition $G_I = 1/G$ where $G = \Sigma/\Delta$ is the gain of the given graph.

The assumption is now made that the numerator and denominator of the system gain is linear in the parameter of interest (k_1). That is

$$\Sigma = \Sigma^o + k_1 \Sigma^1 \tag{7}$$

$$\Delta = \Delta^o + k_1 \Delta^1 \tag{8}$$

where neither Σ^1 or Δ^1 contain k_1. Then it follows that

$$G = \frac{\Sigma}{\Delta} = \frac{\Sigma^o}{\Delta^o} \frac{\Sigma/\Sigma^o}{\Delta/\Delta^o} \tag{9}$$

$$S^G_{k_1} = \frac{k_1}{G}\frac{dG}{dk_1} = \frac{\Delta^o}{\Delta} - \frac{\Sigma^o}{\Sigma} \tag{10}$$

where $S^G_{k_1}$ is the "sensitivity of the system function G to variations in k_1 ".

If it is further assumed that the inverted graph has numerator and denominator linear in k_1 , that is

$$\Sigma_I = \Sigma_I^o + k_1 \Sigma_1^1 \tag{11}$$

$$\Delta_I = \Delta_I^o + k_1 \Delta_1^1 \tag{12}$$

then

$$G_I = \frac{\Sigma_I}{\Delta_I} = \frac{\Sigma_I^o}{\Delta_I^o} \frac{\Sigma_I/\Sigma_I^o}{\Delta_I/\Delta_I^0} \tag{13}$$

$$S^{G_I}_{k_I} = \frac{\Delta_I^o}{\Delta_I} - \frac{\Sigma_I^o}{\Sigma_I} \tag{14}$$

A word about this latter assumption is in order. As long as the inverted path does not contain k_1, the inverted graph will be linear in k_1 if the original graph possessed this property. We also tacitly assume later that Σ^o, Δ^o, Σ_I^o, and Δ_I^o are non-zero. In the language of feedback systems this means there exists a non-zero, finite "leakage" path by passing the parameter k_1. Thus F_{k_i}' is finite. Since $G = 1/G_I$, equations (9) and (13) yield

$$\frac{\Sigma^o}{\Delta^o} \frac{\Sigma/\Sigma^o}{\Delta/\Delta^o} = \frac{\Delta_I^o}{\Sigma_I^o} \frac{\Delta_I/\Delta_I^o}{\Sigma_I/\Sigma_I^o} \tag{15}$$

But $G^o = \Sigma^o/\Delta^o$ must equal $1/G_I^o = \Delta_I^o/\Sigma_I^o$. Thus

$$\frac{\Sigma/\Sigma^o}{\Delta/\Delta^o} = \frac{\Delta_I/\Delta_I^o}{\Sigma_I/\Sigma_I^o} \tag{16}$$

Since $S^G_{k_1} = -S^{G_I}_{k_1}$, equations (10) and (14) yield

$$\frac{\Delta^o}{\Delta} - \frac{\Sigma^o}{\Sigma} = \frac{\Sigma_I^o}{\Sigma_I} - \frac{\Delta_I^o}{\Delta_I} \tag{17}$$

Equations (16) and (17) are independent. To prove this, make the following definitions for algebraic simplicity.

$$a = \frac{\Delta^o}{\Delta} \qquad a' = \frac{\Sigma_I^o}{\Sigma_I} \tag{18}$$

$$b = \frac{\Sigma^o}{\Sigma} \qquad b' = \frac{\Delta_I^o}{\Delta_I} \tag{19}$$

Then from (17) we obtain

$$b = a + b' - a'$$

$$b/a = 1 + b'/a - a'/a \tag{20}$$

But from (16)

$$b/a = b'/a' \tag{21}$$

Since (20) and (21) are not identical, in general, (16) cannot be derived from (17), and conversely; hence these two equations are independent. To satisfy the additional constraint on b/a, it is most natural to let a=a'. This requires that b=b'. By comparing equations (18) with (1) and (5), and (19) with (2) and (6) we see

$$F_{k_1} = \frac{\Delta}{\Delta^o} = \frac{\Sigma_I}{\Sigma_I^o} \qquad (22)$$

$$F'_{k_1} = \frac{\Sigma}{\Sigma^o} = \frac{\Delta_I}{\Delta_I^o} \qquad (23)$$

Note that this derivation requires only that Δ be linear in k_1, and that F_{k_1}' exist, that is to say $\Sigma^o \neq 0$. Hence equations (22) and (23) can be used to determine return differences even though the parameter of interest appears more than once in the SFG. In the event that the parameter appears only once in the graph, these quantities agree with the usual definitions. Also note that equation (9) becomes

$$G = G^o \frac{F_{k_1}'}{F_{k_1}} \qquad (24)$$

which is Blackman's Equation for one parameter and equation (10) becomes

$$S_{k_1}^{G_1} = \frac{1}{F_{k_1}} - \frac{1}{F_{k_1}'} \qquad (25)$$

whick is Bode's expression for sensitivity.

Since the physical interpretation of return signals has no basis if the parameter exists in more than one branch of the SFG, the name "multi-return difference" and "multi-null return difference" is suggested for the mathematical results (22) and (23).

IV EXTENSION TO THE MULTIPARAMETER CASE

The extension is quite straight-forward. Let $G^{(k_1,k_2,\ldots,k_n)}$ be the system gain. Let $G^{(0,0,\ldots,k_{i+1},k_{i+2},\ldots,k_n)}$ be the system gain when $k_1,k_2,..k_i$ have been set to zero. Obviously

$$G^{(k_1k_2,\ldots k_n)} =$$

$$G^{(0,0,\ldots,0)}\,\frac{G^{(0,0,\ldots,k_n)}}{G^{(0,0,\ldots,0)}}\cdot\frac{G^{(0,0,\ldots,k_{n-1},k_n)}}{G^{(0,0,\ldots 0,k_n)}}\cdot\frac{G^{(k_1,k_2,\ldots k_n)}}{G^{(0,k_2,\ldots k_n)}} \tag{26}$$

But by (24)

$$\frac{G^{(0,0,\ldots,k_i,k_{i+1},\ldots k_n)}}{G^{(0,0,\ldots,0,k_{i+1},\ldots k_n)}} = \frac{F^{(i-1)'}_{ki}}{F^{(i-1)}_{k_i}} \tag{27}$$

where $F^{(i-1)}_{k_i}$ is the return difference, with respect to parameter k_i, of the system which results from setting $k_1,k_2,\ldots k_{i-1}$ equal to zero. Substituting (27) into (26) yields the generalized Blackman's equation for n parameters

$$G^{(k_1,k_2,\ldots k_n)} = G^{(0,0,\ldots 0)}\prod_{i=1}^{n}\frac{F^{(i-1)'}_{k_i}}{F^{(i-1)}_{k_i}} \tag{28}$$

The gain can also be written as

$$G^{(k_1,k_2,\ldots k_n)} = \frac{\Sigma^{(k_1,k_2,\ldots k_n)}}{\Delta^{(k_1,k_2,\ldots k_n)}}$$

$$= \frac{\Sigma^{(0,0,\ldots 0)}}{\Delta^{(0,0,\ldots 0)}}\,\frac{\Sigma^{(0,0,k_n)}}{\Sigma^{(0,0,\ldots 0)}}\cdots\frac{\Sigma^{(k_1,k_2,\ldots,k_n)}}{\Sigma^{(0,k_2,\ldots k_n)}}$$

$$\frac{\Delta^{(0,0,\ldots 0)}}{\Delta^{(0,0,\ldots k_n)}}\cdot\frac{\Delta^{(0,0,\ldots k_n)}}{\Delta^{(0,0,\ldots k_{n-1},\,k_n)}}\cdots\frac{\Delta^{(0,k_2,..k_n)}}{\Delta^{(k_1,k_2\ldots,k_n)}} \tag{29}$$

Similar expressions can be written for the gain of the inverted graph. It is obvious that we seek to identify

$$F_{k_i}^{(i-1)} = \frac{\Delta^{(0,0,\ldots,k_{i+1},\ldots,k_n)}}{\Delta^{(0,0,\ldots,0,k_{i+1},\ldots k_n)}} \qquad (30)$$

$$F_{k_i}^{(i-1)'} = \frac{\Sigma^{(0,0,\ldots,k_i,k_{i+1},\ldots k_n)}}{\Sigma^{(0,0,\ldots 0,k_{i+1},\ldots,k_n)}} \qquad (31)$$

To do this consider the sensitivity of $G^{(0,0,\ldots,k_i,k_{i+1},\ldots k_n)}$ to variations in k_i. This is exactly the same as the one parameter problem. In fact, it can be assumed that the system discussed in section III was actually derived from a more complicated system by setting k-1 parameters to zero. By repeating the same arguments it is seen that the identifications (30) and (31) can be made. Provided that a leakage path exists which contains none of the $k_i (i=1,2,\ldots,n)$ then it is possible to further identity.

$$F_{k_i} = \frac{\Sigma_I(0,0,\ldots,k_i k_{i+1},\ldots,k_n)}{\Sigma_I(0,0,\ldots,0,k_{i+1},\ldots,k_n)} \qquad (32)$$

$$F_{k_i}' = \frac{\Delta_I(0,0,\ldots,k_i,k_{i+1},\ldots,k_n)}{\Delta_I(0,0,\ldots,0,k_{i+1},\ldots,k_n)} \qquad (33)$$

Equations (30) and (32) are the "multi-return differences" for the n parameter case, and equations (31) and (33) are the "multi-null return differences" for the n parameter case.

V MULTI-RETURN DIFFERENCES APPLIED TO GAIN AND SENSITIVITY CALCULATIONS

Since it is usually easier to determine the numerator of a SFG than it is to determine the denominator, the results presented above provide a convenient method to determine the gain and sensitivity. The equation for gain is

$$G = G^o \frac{F_k'}{F_k} = G^o \frac{\Sigma/\Sigma^o}{\Sigma_I/\Sigma_I^o} \qquad (34)$$

in the single parameter case and

$$G = G^{(0,0,\ldots 0)} \prod_{i=1}^{n} \left(\frac{\Sigma^{(0,0,k_i k_{i+1},\ldots k_n)}}{\Sigma^{(0,0,\ldots 0,k_{i+1}\ldots k_n)}} \cdot \frac{\Sigma_I^{(0,0,\ldots k_{i+1},\ldots,k_n)}}{\Sigma_I^{(0,0,\ldots k_i,k_{i+1}\ldots k_n)}} \right) \qquad (35)$$

$$G = G^{(0,0,\ldots 0)} \frac{\Sigma^{(k_1,k_2,\ldots,k_n)}}{\Sigma^{(0,0,\ldots,0)}} \frac{\Sigma_I^{(0,0,\ldots,0)}}{\Sigma_I^{(k_1,k_2,\ldots k_n)}} \qquad (36)$$

in the multiparameter case. Sensitivity can only be determined for one parameter at a time, but having determined Σ and Σ_I, the various sensitivities are easily determined from

$$S_{k_i}^{G} = \frac{\Sigma_I^{(k_1,k_2,\ldots k_{i-1},0,k_{i+1}\ldots k_n)}}{\Sigma_I^{(k_1,k_2\ldots k_i,k_{i+1},\ldots k_n)}} - \frac{\Sigma^{(k_1,k_2,\ldots,k_{i-1},0,k_{i+1},\ldots,k_n)}}{\Sigma^{(k_1,k_2,\ldots,k_i,k_{i+1},\ldots,k_n)}} \qquad (37)$$

The procedure for applying equation (36) is as follows. It is most convenient to let $G^{(0,0,\ldots 0)}$ be as uncomplicated as possible. Ideally, one would like to pick parameters $k_1,k_2,\ldots k_n$ so that there will be no loops left in the graph when these parameters are set to zero. However, some of these parameters may also **appear** in the leakage path to be inverted(for constructing the inverted graph), so one approaches the ideal case as nearly as possible. Having picked the parameters of interest, the leakage path is inverted. Then $G^{(0,0,\ldots 0)}$, Σ, and Σ_I are evaluated. Then $\Sigma^{(0,0,\ldots 0)}$ is found directly from $\Sigma (= \Sigma^{(k_1,k_2,,,,k_n)})$ and $\Sigma_I^{(0,0,\ldots,0)}$ is found from $\Sigma_I (= \Sigma^{(k_1,k_2,\ldots k_n)})$. These values can be substituted into (36) to yield the gain. In similar fashion the various sensitivities can be found.

It should be noted that (36) is of the form

$$G = G_{no} \frac{F'_{nk}}{F_{nk}}$$

where

$$G_{no} = G^{(0,0,\ldots 0)}$$

$$F'_{nk} = \frac{\Sigma^{(k_1,k_2,\ldots,k_n)}}{\Sigma^{(0,0,\ldots 0)}}$$

$$F_{nk} = \frac{\Sigma_I^{(k_1,k_2,\ldots,k_n)}}{\Sigma^{(0,0,\ldots 0)}}$$

This form agrees with the generalized Blackman's Equation derived by Sun and Reis$^{(10)}$. Thus, in the event that the various parameters appear only once in the graph, the results of this paper agree with those of Sun and Reis.

VI EXAMPLE

Consider the transistor amplifier shown in figure one. A SFG for this network is shown in figure two. Note that setting R_e, h_{22} and h_r equal to zero results in a graph having no loops. The leakage path contains none of these parameters, and it is known "a priori" that the system determinant is linear in the system parameters. Inverting the leakage path e_1, i_1, i_2, e_2 results in the graph shown in figure three. System gain G (= $G^{(R_e, h_r, h_{22})}$) is then found from

$$G = G^{(0,0,0)} \frac{\Sigma}{\Sigma^{(0,0,0)}} \frac{\Sigma_I^{(0,0,0)}}{\Sigma_I} \tag{38}$$

By inspection

$$G^{(0,0,0)} = \frac{-\beta R_L}{h_i} \tag{39}$$

$$\Sigma = \Sigma^{(R_e, h_r, h_{22})} = \frac{-\beta R_2}{h_i} + \frac{R_e R_L h_{22}}{h_i} \tag{40}$$

$$\Sigma_I = \Sigma_I^{(R_e, h_r, h_{22})} = h_r - \frac{-h_{22} h_i}{\beta} - \frac{h_i}{\beta R_L} - \frac{R_e}{\beta R_L} - \frac{h_r R_e}{\beta R_L} + \frac{h_r R_e}{R_L}$$

$$- \frac{R_e}{R_L} - \frac{R_e h_{22} h_i}{\beta R_L} - \frac{h_{22} R_e}{\beta} \tag{41}$$

From (40) and (41)

$$\Sigma^{(0,0,0)} = - \frac{\beta R_L}{h_i} \tag{42}$$

$$\Sigma_I^{(0,0,0)} = - \frac{h_i}{\beta R_L} \tag{43}$$

Substituting (39) through (43) into (38) and subsequent simplification yields

$$G = \frac{R_e R_L h_{22} - \beta R_L}{R_e \left[1 + \beta - h_r + h_i h_{22} - h_r \beta + h_{22} R_L\right] + R_L (h_i h_{22} - h_r \quad) + h_i} \tag{44}$$

The sensitivity of G to R_e is found from

$$S_{R_e}^{G} = \frac{\Sigma^{(0, h_r, h_{22})}}{\Sigma} - \frac{\Sigma_I^{(0, h_r h_{22})}}{\Sigma_I} \tag{45}$$

which corresponds to (37). The result is

$$S_{R_e}^{G} = \frac{-\beta R_L}{R_e R_L h_{22} - \beta R_L} - \frac{h_r \beta R_L - h_i - h_{22} h_i R_L}{h_r \beta R_L - h_i - h_{22} h_i R_L - R_e \left[1 + h_{22} h_i + \beta + h_{22} R_L - h_r \beta - h_r\right]} \tag{46}$$

In a similar fashion

$$S_{h_{22}}^{G} = \frac{-\beta R_L}{R_e R_L h_{22} - \beta R_L} - \frac{h_r \beta R_L - h_i - R_e \left[1 + \beta - h_r - h_r \beta\right]}{h_r \beta R_L - h_i - h_{22} h_i R_L - R_e \left[1 + h_{22} h_i + \beta + h_{22} R_L - h_r \beta - h_r\right]}$$

$$S_{h_r}^{G} = 1 - \frac{-h_i - h_{22} h_i R_L - R_e \left[1 + h_{22} h_i + \beta + h_{22} R_L\right]}{h_r \beta R_L - h_i - h_{22} h_i R_L - R_e \left[1 + h_{22} h_i + \beta + h_{22} R_L - h_r \beta - h_r\right]}$$

To determine S_{β}^{G}, the inverted graph is not suitable since β is in the inverted path. This could be remedied by inverting the path along the nodes $e_1, i_1, V_e V_2, i_2 e_2$, thus putting β in a leakage path. Note that neither $S^G_{h_i}$ or $S^G_{R_L}$ can be found by the present technique since all forward paths include these parameters. Hence the null return differences do not exist.

VII SUMMARY

The return differences are fundamental quantities used to analyze feedback systems. New methods for evaluating these return differences have recently been reported(8,9,10,11). These results were limited to the case in which the parameters of interest appeared as the transmittances of a single branch of the SFG. This paper removes this limitation, subject to the rather mild restriction that the system determinant be linear in the parameters, and the vanishing of the parameters does not cause the system gain to vanish. This additional freedom results in particularly simple techniques for obtaining system gain and sensitivity.

REFERENCES

1. H. W. Bode, Network Analysis and Feedback Amplifier Design. New York: D. Van Nostrand, pp. 33-102, 1945.

2. S. J. Mason, "Feedback Theory -- Some Properties of Signal Flow Graphs," Proc. IRE (Sept. 1953), pp. 1144-1156.

3. S. J. Mason, "Feedback Theory -- Further Properties of Signal Flow Graphs," Proc. IRE (July 1956), pp. 920-926.

4. J. G. Truxal, Control System Synthesis. New York: McGraw-Hill, p. 114, 1955.

5. L. P. A. Robichaud, M. Boisvert, J. Robert, Signal Flow Graphs and Applications. Englewood Cliffs, N. J.: Prentice-Hall, p. 31, 1962.

6. I. M. Horowitz, Synthesis of Feedback Systems. New York: Academic Press, p. 70, 1963.

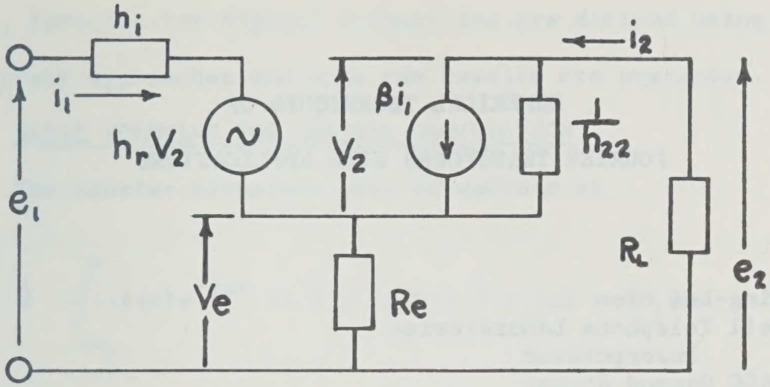
7. B. C. Kuo, Automatic Control Systems. Englewood Cliffs, N. J.: Prentice-Hall, p. 73, 1962.

8. H. H. Sun and G. C. Reis, "Feedback System Analysis by Four Node Graphs," Proc. of the First Allerton Conference on Circuit and System Theory (Nov. 1963), pp. 764-778.

9. H. H. Sun and G. C. Reis, "General Methods of Finding the Null Return Difference From a Signal Flow Graph," Proc. of the Sixth Midwest Symposium on Circuit Theory (May 1963), pp. H1-H20.

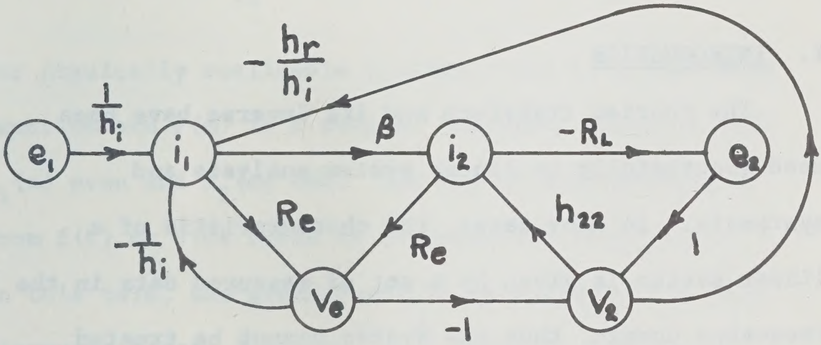
10. G. C. Reis, General Method for Determining Null Return Ratio. M. S. Thesis, Drexel Institute of Technology Library, June 1962.

11. H. H. Sun and G. C. Reis, "Return Difference and Null Return Difference of a Feedback System," IEEE Trans. of PGAC (Jan. 1964), pp. 109-110.



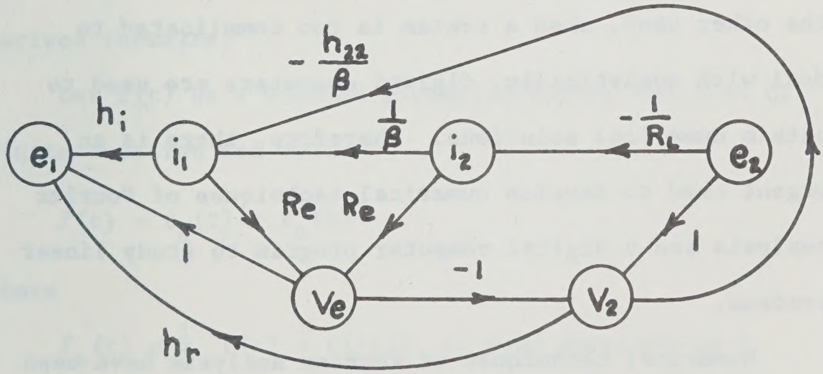
h_i
i_2
i_1
βi_1
$\frac{1}{h_{22}}$
$h_r V_2$
V_2
e_1
R_L
e_2
V_e
Re

Figure 1



$-\frac{h_r}{h_i}$
$\frac{1}{h_i}$
β
$-R_L$
e_1
i_1
i_2
e_2
Re
h_{22}
$-\frac{1}{h_i}$
Re
1
V_e
V_2
-1

Figure 2



$-\frac{h_{22}}{\beta}$
h_i
$\frac{1}{\beta}$
$-\frac{1}{R_L}$
e_1
i_1
i_2
e_2
Re
Re
1
1
-1
V_e
V_2
h_r

Figure 3

NUMERICAL TECHNIQUES OF FOURIER TRANSFORMS WITH APPLICATIONS

Ming-Lei Liou
Bell Telephone Laboratories
Incorporated
1600 Osgood Street
North Andover, Massachusetts

I. INTRODUCTION

The Fourier transform and its inverse have been used successfully in linear system analysis and synthesis. In many cases, the characteristic of a linear system is given by a set of measured data in the frequency domain, thus the system cannot be treated analytically and numerical methods must be used. On the other hand, when a system is too complicated to deal with analytically, digital computers are used to obtain numerical solutions. Therefore, there is an urgent need to develop numerical techniques of Fourier analysis and a digital computer program to study linear systems.

Numerical techniques of Fourier analysis have been considered by Guillemin(1) and Papoulis.(2) In this talk, formulas for digital computation are derived using different approaches and some new results are presented.

II. BASIC FORMULAS FOR DIGITAL COMPUTATION

The Fourier transform pair is defined as

$$F(\omega) \triangleq \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt \triangleq \mathcal{F}[f(t)] \triangleq F_1(\omega) + jF_2(\omega)$$

$$f(t) \triangleq \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)e^{j\omega t}\,d\omega \triangleq \mathcal{F}^{-1}\,[F(\omega)].$$

For physically realizable systems, $f(t)$ is a real time function and $F(\omega)$ is a complex frequency function with $F_1(\omega)$ even and $F_2(\omega)$ odd. One way of evaluating $F(\omega)$ from $f(t)$ or vice versa is by numerical integration. In this talk, the given function is approximated by a piecewise linear function$^{(3),(4)}$ and its Fourier or inverse Fourier transform is evaluated from analytically derived formulas.

Let $f(t)$ be a bounded piecewise linear function of finite duration and write

$$f(t) = f_e(t) + f_o(t)$$

where

$$f_e(t) = \tfrac{1}{2}\,[f(t) + f(-t)], \text{ an even function of } t$$

$$f_o(t) = \tfrac{1}{2}\,[f(t) - f(-t)], \text{ an odd function of } t.$$

From the Fourier transform theorem,

$$\mathcal{F}[\frac{df}{dt}] = \mathcal{F}[\frac{df_e}{dt} + \frac{df_o}{dt}] = j\omega F(\omega). \qquad (1)$$

Since $f_e(t)$ is an even piecewise linear function, its derivative consists of a sum of two parts, $a_o(t)$ and $b_o(t)$, where $a_o(t)$ is a set of delta functions distributed oddly and $b_o(t)$ is an odd piecewise constant function. Similarly, the derivative of $f_o(t)$ consists of a sum of two parts, $a_e(t)$ and $b_e(t)$, where $a_e(t)$ is a set of delta functions distributed evenly and $b_e(t)$ is an even piecewise constant function. From Equation (1)

$$\mathcal{F}[b_e(t) + b_o(t)] = j\omega F(\omega) - \mathcal{F}[a_e(t) + a_o(t)]$$

and hence

$$\mathcal{F}[\frac{db_e}{dt} + \frac{db_o}{dt}] = j\omega\left\{ j\omega F(\omega) - \mathcal{F}[a_e(t) + a_o(t)] \right\}. \qquad (2)$$

Since $b_e(t)$ is an even piecewise constant function, its derivative consists of a set of delta functions distributed oddly. Similarly, the derivative of $b_o(t)$ is a set of delta functions distributed evenly.

Let

$$a_e(t) = \sum_{i=-N}^{N} A_i\ \delta(t-t_i);\ A_i = A_{-i},\ t_i = -t_{-i}$$

$$a_o(t) = \sum_{i=-N}^{N} B_i\ \delta(t-t_i);\ B_i = -B_{-i},\ t_i = -t_{-i}$$

$$\frac{db_o}{dt} = \sum_{i=-N}^{N} C_i\ \delta(t-t_i);\ C_i = C_{-i},\ t_i = -t_{-i}$$

$$\frac{db_e}{dt} = \sum_{i=-N}^{N} D_i\ \delta(t-t_i);\ D_i = -D_{-i},\ t_i = -t_{-i}\ .$$

Substituting the above equations into Equation (2), the real and imaginary parts of $F(\omega)$, respectively $F_1(\omega)$ and $F_2(\omega)$, are

$$F_1(\omega) = \frac{-1}{\omega^2}\left[C_o + 2\sum_{i=1}^{N} C_i\ \cos\omega t_i\right] - \frac{2}{\omega}\sum_{i=1}^{N} B_i\ \sin\omega t_i \tag{3}$$

$$F_2(\omega) = \frac{2}{\omega^2}\sum_{i=1}^{N} D_i\ \sin\omega t_i - \frac{1}{\omega}\left[A_o + 2\sum_{i=1}^{N} A_i\ \cos\omega t_i\right]. \tag{4}$$

At $\omega = o$, by the definition of the inverse Fourier transform

$$F(o) = \int_{-\infty}^{\infty} f(t)\ dt = 2\int_{o}^{\infty} f_e(t)\ dt\ . \tag{5}$$

Equations (3), (4), and (5) are the three basic equations for digital computation. Since $F_1(\omega)$ and

$F_2(\omega)$ are respectively even and odd functions, they are evaluated only for $\omega \geq o$.

The coefficients A_i, B_i, C_i and D_i in Equations (3) and (4) are given as follows:

$$\left.\begin{aligned} A_i &= f_o(t_i^+) - f_o(t_i^-),\ i = 1,2,\ldots,N \\ A_o &= 2f_o(t_o^+) \end{aligned}\right\} \quad (6)$$

$$B_i = f_e(t_i^+) - f_e(t_i^-),\ i = 1,2,\ldots,N \quad (7)$$

$$\left.\begin{aligned} C_i &= \frac{f_e(t_{i+1}^-) - f_e(t_i^+)}{t_{i+1} - t_i} - \frac{f_e(t_i^-) - f_e(t_{i-1}^+)}{t_i - t_{i-1}}, \\ &\qquad i=1,2,\ldots,N \\ C_o &= \frac{2}{t_1}\,[f_e(t_1^-) - f_e(t_o^+)] \end{aligned}\right\} \quad (8)$$

$$\left.\begin{aligned} D_i &= \frac{f_o(t_{i+1}^-) - f_o(t_i^+)}{t_{i+1} - t_i} - \frac{f_o(t_i^-) - f_o(t_{i-1}^+)}{t_i - t_{i-1}}, \\ &\qquad i=1,2,\ldots,N \end{aligned}\right\} \quad (9)$$

where

$$t_{i+1} > t_i \text{ and } t_o = 0 \quad (10)$$

$$\left.\begin{aligned} f_e(t_i^+) = f_e(t_{i+1}^-) = f_o(t_i^+) = f_o(t_{i+1}^-) = 0 \\ \text{for } i \geq N \end{aligned}\right\}, \quad (11)$$

The range of ω needed for computation in Equation

(3) is either given beforehand or determined as follows:

Let

$$2|C_o| + \sum_{i=1}^{N} |C_i| = k_1$$

$$\sum_{i=1}^{N} |B_i| = k_2$$

$$\int_o^{\infty} |f_e(t)|\, dt = I_1.$$

Then, the function $F_1(\omega)$ is bounded by $2I_1$ and $2\left\{\frac{k_1}{\omega^2} + \frac{k_2}{|\omega|}\right\}$ as shown in Figure 1.

Thus, the range of ω is from 0 to ω_e,

$$\omega_e = \frac{1}{2\epsilon I_1}\left[k_2 + \sqrt{k_2^{\,2} + 4\epsilon I_1 k_1}\right]$$

where ϵ (< 1) is an arbitrary constant which determines the value of ω_e such that $|F_1(\omega)|$ is less than certain percentage of max. $|F_1(\omega)|$ for $\omega \geq \omega_e$.

Similarly, for Equation (4)

$$2|A_o| + \sum_{i=1}^{N} |A_i| = k_3$$

$$\sum_{i=1}^{N} |D_i| = k_4$$

$$\int_{o}^{\infty} |f_o(t)|\, dt = I_2$$

and

$$\omega_o = \frac{1}{2\epsilon\, I_2} \left[k_3 + \sqrt{k_3^{\,2} + 4\epsilon\, I_2\, k_4} \right]$$

where ϵ is defined as before. The value of ω_o is compared with the value of ω_e and the larger one is chosen for the range of computation.

As an example, consider the piecewise linear time function shown in Figure 2. The even and odd functions and their derivatives are, respectively, shown in Figures 3 and 4. The coefficients are evaluated using Equations (6) to (11) and given in Table I. From Equations (3), (4) and (5),

$$F_1(\omega) = \frac{1}{2\omega^2} [\cos \omega - \cos 3\omega] + \frac{\sin 4\omega}{\omega}$$

$$F_2(\omega) = \frac{1}{\omega^2} [\sin\omega - \sin 3\omega] + \frac{2\cos 4\omega}{\omega}$$

$$F(o) = 6,$$

The solution is exact because the given function is piecewise linear.

Table I - Coefficients of Delta Functions of f(t)

i	0	1	2	3
t_i	0	1	3	4
A_i	0	0	0	-1
B_i	-	0	0	-0.5
C_i	0	-0.25	0.25	0
D_i	-	0.5	-0.5	0

Same formulas can be used to obtain the inverse Fourier transform of a given complex frequency function, $F(\omega) = F_1(\omega) + jF_2(\omega)$. The procedure is as follows:

(1) Replace the argument, ω, of $F(\omega)$ by t with $F_1(t)$ even and $F_2(t)$ odd

(2) evaluate $\mathcal{F}[F_1(t)] = 2\pi f_e(\omega)$ and $\mathcal{F}[-jF_2(t)] = 2\pi f_o(\omega)$ using the formulas for the Fourier transform [i.e., Equations (3), (4), and (5)]

(3) divide the sum of $\mathcal{F}[F_1(t)]$ and $\mathcal{F}[-jF_2(t)]$ by 2π and replace its argument, ω, by t.

III. MODULATED FUNCTIONS AND BANDPASS SYSTEMS

Consider the modulated function

$$f(t) = f_m(t) \cos (\omega_o t + \theta) \tag{12}$$

where $f_m(t)$ is a slowly varying arbitrary modulating function.

A direct application of the method presented in the last section requires a lengthy calculation and perhaps involves an appreciable amount of error. In order to reduce the error, the cosine function in Equation (12) should be treated analytically.

It can be shown that the Fourier transform of Equation (12) is

$$\mathcal{F}[f_m(t)\cos(\omega_o t+\theta)] = \frac{1}{2}e^{j\omega\frac{\theta}{\omega_o}}\Big\{[F_{m1}(\omega+\omega_o)+F_{m1}(\omega-\omega_o)]$$
$$+j\,[F_{m2}(\omega+\omega_o)+F_{m2}(\omega-\omega_o)]\Big\}$$

where $F_{m1}(\omega)$ and $F_{m2}(\omega)$ are respectively the Fourier transforms of the even and odd parts of $f_m(t-\frac{\theta}{\omega_o})$.

The impulse response of a bandpass system is an oscillatory function. The envelope of the oscillatory function is often of the greatest interest. A direct application of the numerical method developed previously would give a complete oscillatory function which is time consuming and is not necessary.

Consider the bandpass system shown in Figure 5 where the amplitude and phase are respectively even and odd functions of frequency.

Let

$$H_r(\omega) = H(\omega)\ u(\omega)$$

$$H_\ell(\omega) = H(\omega)\ u(-\omega)$$

where $u(\omega)$ is the unit step function of ω.

Define

$$H_p(\omega) \triangleq \frac{1}{2}\,[H_r(\omega+\omega_o) + H_\ell\ (\omega-\omega_o)]$$

$$H_q(\omega) \triangleq \frac{1}{2j}\,[H_\ell\ (\omega+\omega_o) - H_r\ (\omega-\omega_o)]$$

It can be shown that the impulse response of $H(\omega)$ is

$$h(t) = 2h_p(t) \cos\omega_o t + 2h_q(t) \sin \omega_o t$$

where $h_p(t)$ and $h_q(t)$ are respectively the inverse Fourier transforms of $H_p(\omega)$ and $H_q(\omega)$, and are slowly varying envelope functions of $h(t)$.

IV. LINEAR SYSTEM ANALYSIS AND SYNTHESIS

It is advantageous to solve linear system problems in the frequency domain since the input-output relationship is expressed by the following simple algebraic equation

$$F(\omega) = H(\omega)\ G(\omega) \tag{13}$$

where $F(\omega)$, $H(\omega)$, and $G(\omega)$ are respectively the output, system, and input frequency functions.

Any known function expressed in the time domain is first converted to the frequency function. The unknown frequency function is obtained using Equation (13) and its corresponding time function is evaluated by the numerical techniques discussed previously.

The Fourier transform of a step function Ku(t) is

$$\mathcal{F}[Ku(t)] = K\pi\delta(\omega) - j\frac{K}{\omega}.$$

It is very difficult to treat this frequency function numerically. However, in the time domain, the response f(t) of a linear system due to a step input function Ku(t) can be obtained by the numerical integration according to the following equation

$$f(t) = K \int_{-\infty}^{t} h(\tau)d\tau$$

where the impulse response h(t) is evaluated from the system function H(ω) by the numerical method discussed before.

A digital computer program is now available at Bell Telephone Laboratories to study linear systems. The program will analyze or synthesize linear systems with functions expressed in either the time or the frequency domain. At various stages of computation, the results can be (either or) both printed out and plotted by the microfilm plotter. For commonly used input functions such as rectangular pulses, raised cosine, etc., their Fourier transforms are derived analytically and stored in the computer for direct use.

The following example illustrates a typical linear system problem. The input to the system is a rectangular modulated cosine function shown in Figure 6. The system is of the band elimination type shown in Figure 7. The output of the system is shown in Figure 8. Since the system function has infinite frequency range, the product of $H(\omega)$ and $G(\omega)$ has long tail sections. A straightforward application of the inverse Fourier transform techniques to $F(\omega) = H(\omega)\ G(\omega)$ results in errors because of the truncation of the tail sections. However, a band elimination system can be replaced by an all pass system minus a bandpass system, the output function is therefore equal to the input function minus the response of the bandpass system excited by the same input function. The error introduced by the truncation of the tail sections is completely eliminated.

V. OTHER APPLICATIONS

(1) Nonharmonic Analysis of Arbitrary Real Functions; Very often, it is desirable to approximate a real function $f(t)$ by a sum of sinusoids of frequencies which are not necessarily harmonically related.

$$f(t) \cong \sum_{i=1}^{N} a_i \cos\omega_i t + \sum_{i=1}^{N} b_i \sin\omega_i t \qquad (14)$$

The first step of the analysis is to find the values of ω_i, i=1, ω, ..., N. Then, the standard least mean square approximation can be applied to determine the coefficients a_i's and b_i's in Equation (14).

The numerical techniques of Fourier transforms discussed in this talk can be used here to find the magnitude square of $F(\omega) = \mathcal{F}[f(t)]$ which determines the approximate values of ω_i, i=1, 2, ..., N.

(2) Autocorrelation Function and Power Spectral Density: The autocorrelation function and power spectral density are related by the Fourier transform pairs. Since they are real even functions of the respective independent variables, only Equations (3) and (5) are used to evaluate the corresponding transforms.

VI. CONCLUSIONS

Some numerical techniques of Fourier analysis have been presented in this talk. The given time or frequency function is first approximated by linear segments. The corresponding transform is then evaluated from the analytically derived formulas. One set of formulas can be used for evaluating both the Fourier and the inverse Fourier transforms. Ranges needed for computation can be specified in an analytical manner. Various applications such as linear system analysis and synthesis, autocorrelation function and power spectral density, and nonharmonic analysis of arbitrary real functions, have been discussed.

The numerical method presented in this talk is valid only for bounded functions of finite duration. For functions which are not bounded or of infinite duration, some indirect methods must be used. For example, if the input function of a linear system is a step function, the response is obtained by numerical integration.

ACKNOWLEDGMENT

The author wishes to thank Drs. H. Simon and J. J. Lang for their encouragement and assistance in the preparation of this talk.

REFERENCES

(1) E. A. Guillemin, Synthesis of Passive Networks, John Wiley, 1957, Chapter 15.

(2) A. Papoulis, The Fourier Integral and its Applications, McGraw-Hill, 1962, Chapter 4.

(3) R. E. Bellman and S. E. Dreyfus, Applied Dynamic Programming, Princeton University Press, 1962, pp. 334-335.

(4) D. S. Rane, Private Communication

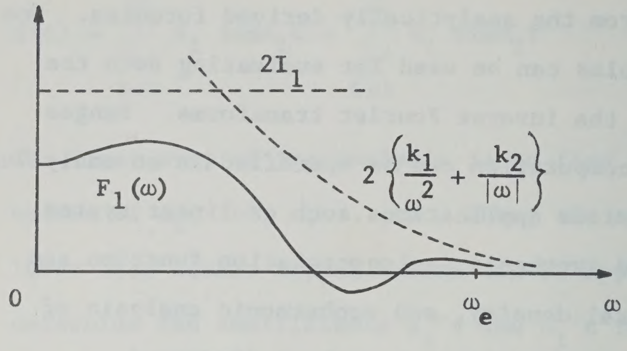


Figure 1 - Bounds of the Function $F_1(\omega)$

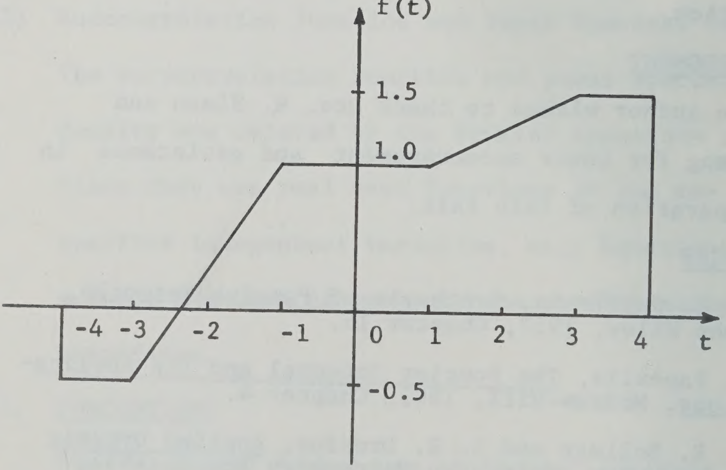


Figure 2 - A Piecewise Linear Time Function

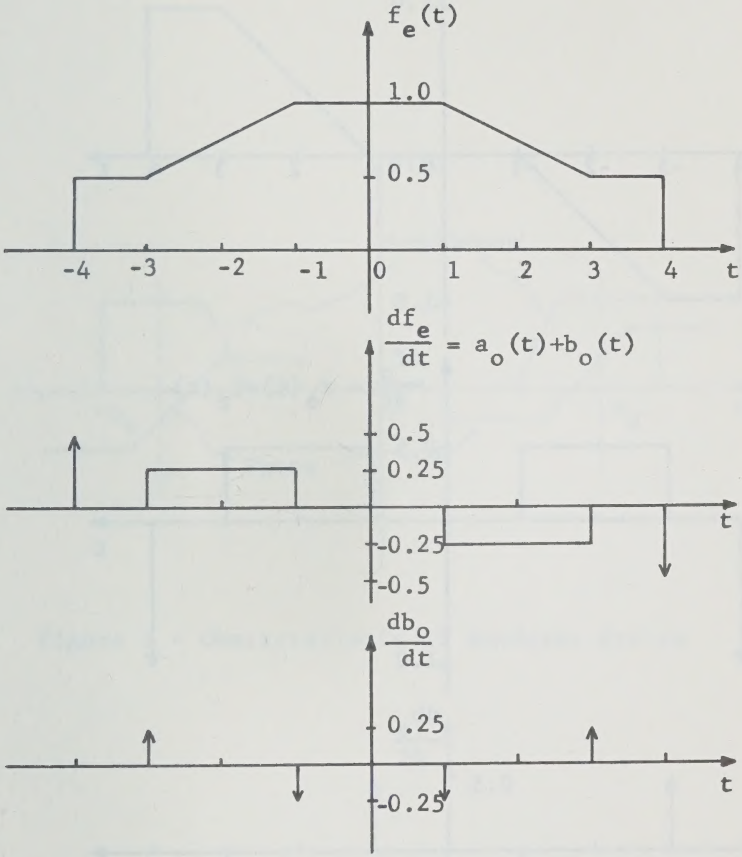


Figure 3 - The Even Time Function and its Derivatives

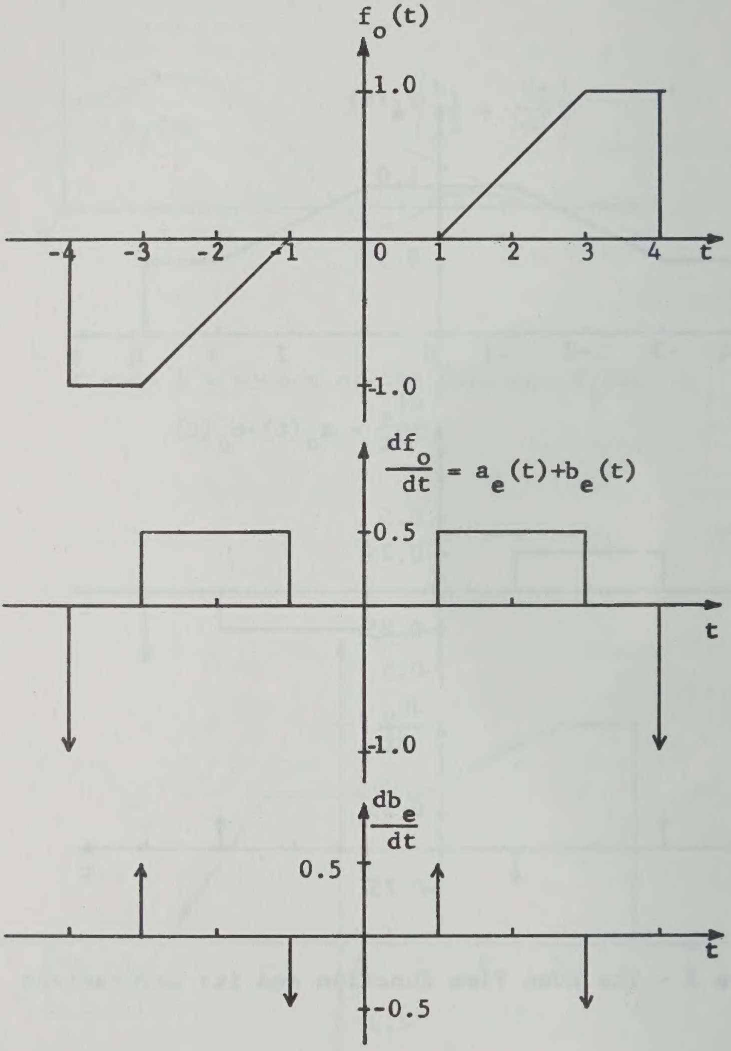


Figure 4 - The Odd Time Function and its Derivatives

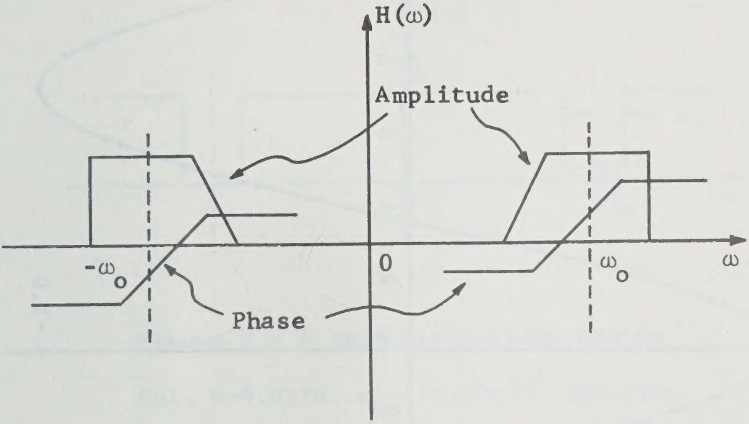

H(ω)
Amplitude
-ω$_o$
0
ω$_o$
ω
Phase

Figure 5 - Characteristic of Bandpass System

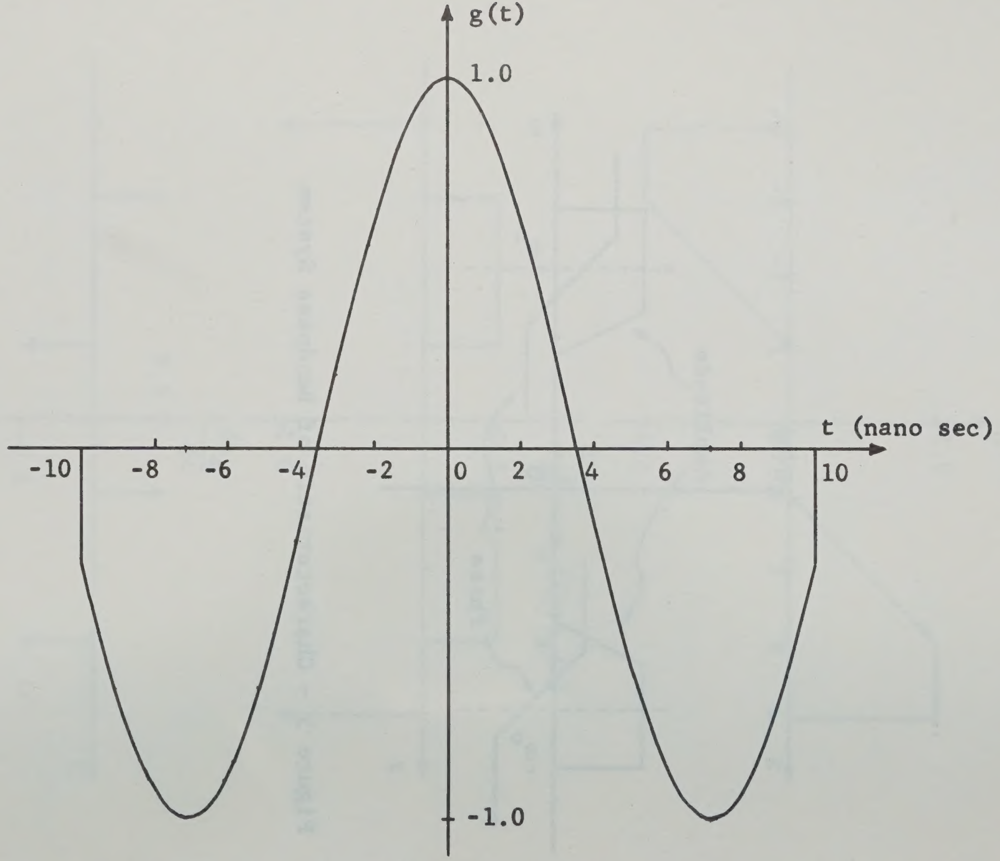


Figure 6 - A Rectangular Modulated Cosine Function

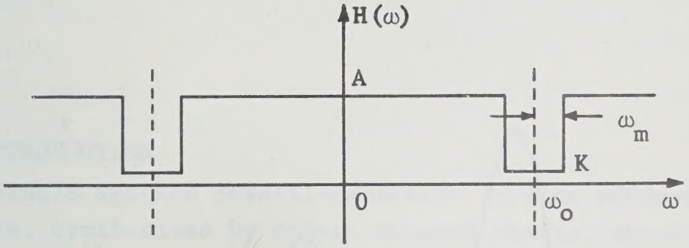


Figure 7 - A Band Elimination System

$A=1$, $K=0.0316$, $\omega_o = 2\pi \times 70 \times 10^6$ rad./sec.

$\omega_m = 2\pi \times 10^7$ rad./sec.

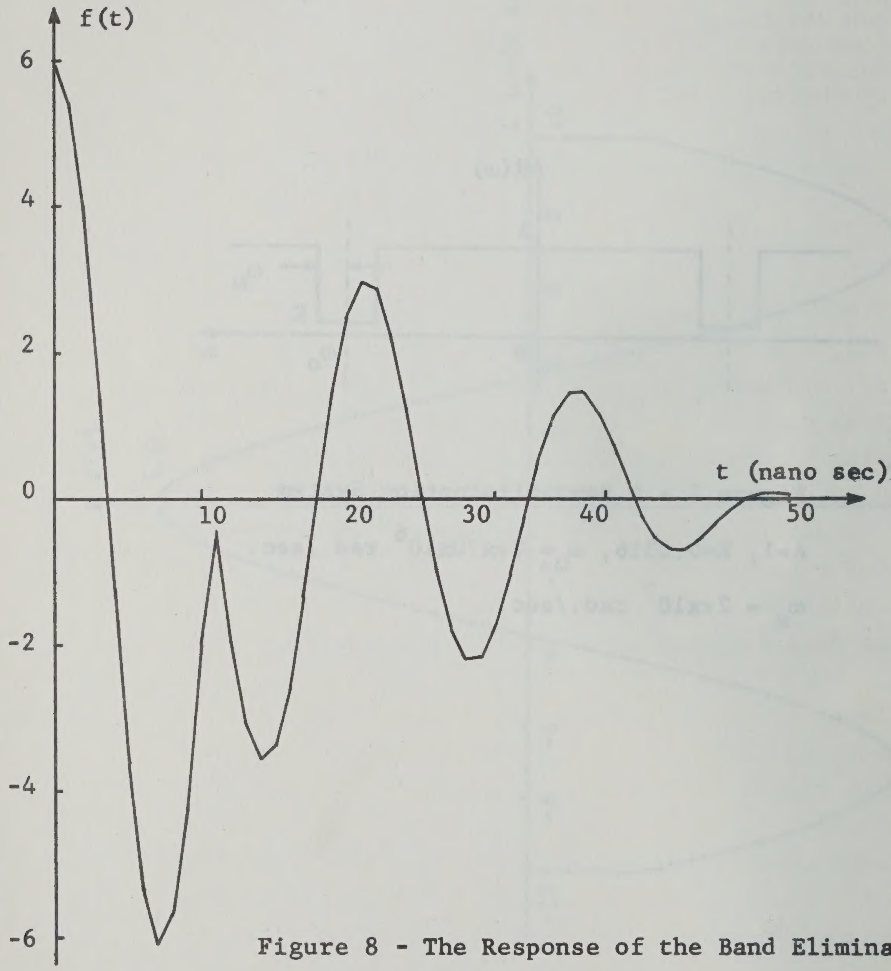


Figure 8 - The Response of the Band Elimination System

SYNTHESIS DESIGN OF CASCADED LATTICE CRYSTAL BANDPASS FILTER

King Yen Cheng
General Electric Company
Communication Products Department
Mountain View Road
Lynchburg, Virginia

I. INTRODUCTION

Single section crystal-capacitor lattice bandpass filters, synthesized by modern network theory, encounter practical difficulties in achieving high values of attenuation in the "stopband", e.g., imperfect balance caused by component tolerance, by extraneous components, etc. In addition to these practical problems, the spurious responses present in the crystals allow certain frequencies to pass unchecked through the filter which further deteriorates the filter performance in its stopband.

In order to avoid the above difficulties, filter engineers may resort to the image parameter method in in which simple 6th order sections are cascaded. Unfortunately, lattice sections can seldom be put in cascade without a mismatch in impedance. This mismatch makes the overall filter response difficult to predict in the beginning, and the whole design requires a cut and try approach. Sometimes it is found necessary to insert an attenuation pad between two consecutive sections in order to minimize impedance mismatch, and the gain of the communication system is deleteriously compromised.

Recently, a method for the synthesis of multiple section lattice filters from low-pass prototypes has been developed.(5) Although this newer method has an advantage over the image parameter design, being free from impedance mismatch between adjacent sections, it permits poles to be located only at zero and infinite frequencies, and frequently it does not yield an efficient design.

The present paper describes a new technique of synthesizing cascaded crystal lattice bandpass filters with poles at finite frequencies, in which each lattice section contains one crystal in each branch.

II. OUTLINE OF THE DESIGN PROCEDURE

As in any synthesis of a lossless network, the design starts with a characteristic function $K(\lambda) = \frac{F(\lambda)}{P(\lambda)}$ such that the

$$\text{Attenuation} = 10 \log (1 + |K(\lambda)|^2) \text{ dbs.}$$

Once $K(\lambda)$ is determined, the transmission function $H(\lambda) = \frac{E(\lambda)}{P(\lambda)}$ can be calculated by the regular method.(1,2) The next step is to realize the network in the form of cascaded lattices as shown in Figure 1.

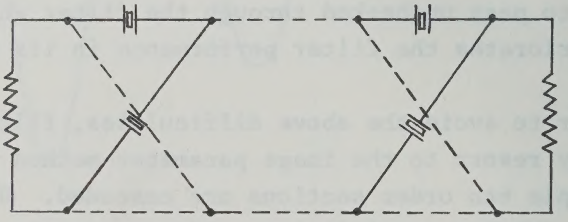


Figure 1

Since each lattice has one crystal in each branch, it can be seen immediately that it is a 6th order symmetrical network possessing two finite poles and two finite zeros on the imaginary axis and two extra zeros on the real axis of the complex frequency plane. If two such sub-networks are in cascade, the overall transmission function H will be given by the expression

$$H = \frac{1}{2}(A + B + C + D)$$

where A B C & D are the four parameters of the chain matrix of the combined network given by:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \times \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix}$$

where A_1 B_1 C_1 D_1 and A_2 B_2 C_2 D_2 are the chain matrices of two sub-networks.

One interesting point to be noted is that if both the sub-networks are symmetrical, that is $A_1 = D_1$ and $A_2 = D_2$, the combined network in general is not symmetrical because,

$$A = A_1A_2 + B_1C_2$$

$$D = D_1D_2 + C_1B_2 = A_1A_2 + C_1B_2$$

and

$$A \neq D \text{ unless } B_1C_2 = C_1B_2.$$

In case $B_1C_2 = C_1B_2$, due to the particular form of the B's and C's, it can be concluded that

$$C_1 = C_2$$

$$B_1 = B_2.$$

Conversely, if we have a symmetrical characteristic function, in general, the network cannot be realized in the form of cascaded symmetrical sub-networks, except, of course, in the form of identical sub-networks.

Based on the previous statement, it can be concluded that a symmetrical characteristic function, $K(\lambda)$, which is to be realized in the form of cascaded lattices, must always have some "left over" elements between symmetrical sub-networks or at the end. Since it is difficult to predict the form of the leftover elements, it would be almost impossible to factor mathematically the chain matrix of the network specified by $K(\lambda)$ into products of chain matrices which in turn define the desired sub-networks. Therefore, the network must be realized in ladder form and transformed section by section into simple lattices.

III. CHOICE OF CHARACTERISTIC FUNCTION K(λ)

It is known that a crystal-capacitor bandpass filter has a characteristic function of the following form:

$$K(\lambda) = \frac{(\lambda^2-k)\prod_{i=1}^{2n}(\lambda^2+\Omega_{o,i})}{h\lambda\prod_{i=1}^{2n}(\lambda^2+\Omega_{\infty i})},$$

where h is a constant depending on the specified maximum attenuation in the passband,

k is a positive parameter

2n is number of finite poles or zeroes

Ω_o's are zeroes on the imaginary axis of the λ plane

and

Ω_∞'s are poles on the imaginary axis of the λ plane.

From the characteristic function K(λ), it can be seen immediately that:

a) There are as many finite zeroes in the passband as there are finite poles in the stop band, and both zeroes and poles occur in pairs.

b) A single pole is found at zero frequency and another at infinite frequency.

c) The numerator of K(λ) is one degree higher than the denominator, and, therefore, the filter must be symmetrical.

The filter which was chosen to be realized is a succession of lattices with one crystal in each branch. Each lattice section will have a finite pole below and a finite pole above the center frequency. It is necessary, therefore, that the number of Ω_∞'s below and above the center frequency of the filter be the same.

The choice of the locations of the Ω_o's and the Ω_∞'s depends on the specifications of the filter. In practice, the passband of the filter is usually very narrow compared to the center frequency. The factor, $\frac{\lambda^2-k}{\lambda}$, in K(λ) has very little effect in the immediate vicinity of the center frequency. For all practical purposes, the zeroes and poles can be assigned as if the filter were asymmetrical and of two orders lower. All known techniques can be applied to design of the filter to meet different requirements. The filter can be designed to have a Cauer response, i.e., equal ripple in both the passband and the stop band. It can also be designed to have inverted Tchebyscheff response, i.e., equal ripple in the stop band and maximally flat in the passband. Felzer's asymmetrical bandpass formula can be used, if the filter is of general parameter design.

Choice of the value of k is rather arbitrary. k may have any value, as long as it is manageable positive number. However, it has been found that for normalized filters, i.e., where the center frequency is referred to as 1, the crystals of the filter will have the highest Cp/Cs ratio when k = 1.

IV. REALIZATION OF THE FILTER IN LADDER FORM

After the characteristic function, $K(\lambda) = \frac{F(\lambda)}{P(\lambda)}$, has been determined, the transmission function $H(\lambda) = \frac{E(\lambda)}{P(\lambda)}$ can be calculated by the regular method.(1,2) Once the $E(\lambda)$ polynominal is known, the filter can be realized in ladder form as shown in Figure 2.

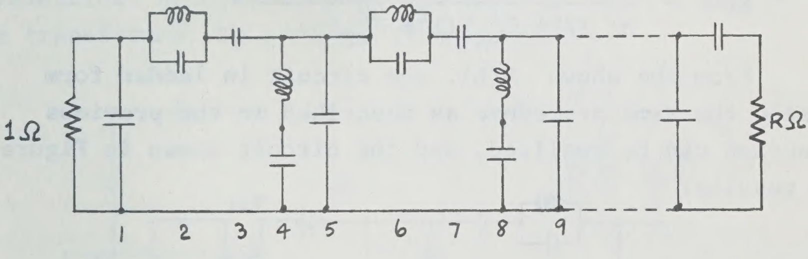


Figure 2

The tank circuits in the series branches define the pole frequencies above the center frequency, while the resonant circuits in the shunt branches define the pole frequencies below the center frequency. In this circuit, it is always advantageous to realize the pair of the two remotest poles first, then realize the pair of next closer poles, and so on until the pair of the closest poles are realized in the last branches of the filter.

V. TRANSFORMATION FROM LADDER INTO CASCADED LATTICES

The filter realized in ladder form as described in the previous section, cannot, in most cases, be implemented by practical crystals. It must be transformed into lattice form, so that it can actually be built. Before the transformation technique is discussed, consider an elementary lattice section as shown in Figure 3.

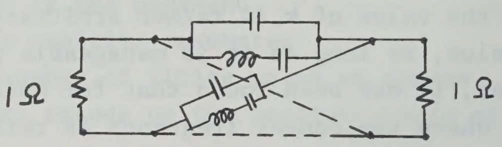


Figure 3

Since, in practice, only lattices which can be transformed into half lattices by using a 1:-1 transformer are of interest, the circuit shown in Figure 3 is assumed to be symmetrical.

The characteristic function, K(λ), of the lattice section is of the form

$$K(\lambda) = \frac{(\lambda^2-k)(\lambda^2+\Omega_{01}^2)(\lambda^2+\Omega_{02}^2)}{h\lambda\ (\lambda^2+\Omega_{\infty 1}^2)(\lambda^2+\Omega_{\infty 2}^2)} .$$

From the above K(λ), the circuit in ladder form using the same procedure as described in the previous section can be realized, and the circuit shown in Figure 4 results.

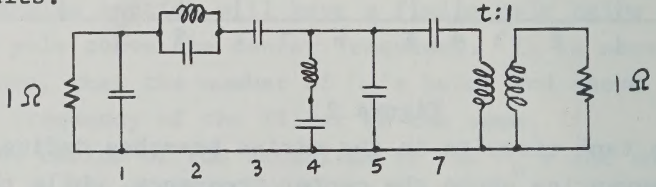


Figure 4

Since both the circuits in Figure 3 and Figure 4 are derived from the same K(λ), they are necessarily equivalent. In other words, given a ladder network in the form

Figure 4, it can always be transformed into a lattice by proper means provided the network is symmetrical.

Now consider the network of Figure 2. The first elements (C_1, L_2, C_2, C_3, L_4, C_4, C_5, C_7) are arranged in a manner similar to that of Figure 4. If these elements can be rearranged, that portion of the ladder can be transformed as a lattice section. The prerequisite condition for the transformation is to make a symmetrical ladder similar to the circuit in Figure 4 out of those elements. In order to accomplish this, the portion of the ladder network in question is redrawn in the form shown in Figure 5.

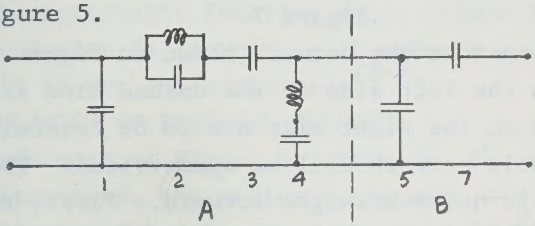


Figure 5

The circuit in Figure 5 can be divided into two parts as shown by the dashed line. Now consider part A of the circuit. In order to make part A a portion of a symmetrical network as shown in Figure 4, part B has to be transformed, by a Norton transformation,

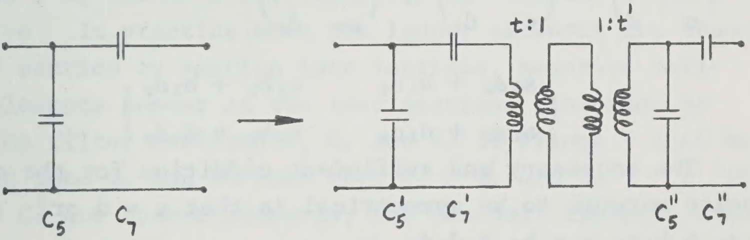


Figure 6

into the circuit shown in Figure 6.

If the values of C_5', C_7' and the transformer, t , can be chosen such that when combined with part A of Figure 5, they will make a symmetrical network, then it can be transformed into a lattice. The residual elements C_5'', C_7'', and the transformer t' will combine with the next portion of the ladder network of Figure 2 for further transformation into a lattice. The same process can be repeated until the entire circuit of Figure 2 is transformed.

VI. DETERMINATION OF C_5', C_7', and t

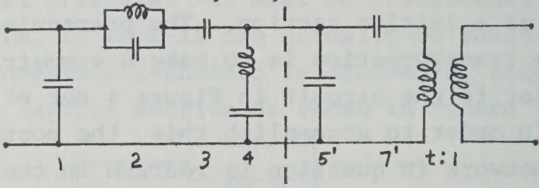


Figure 7

Next consider the network shown in Figure 7. The elements on the left side of the dashed line are known, while those on the right side are to be determined so that the whole network will be symmetrical. The method to be used is quite straight forward. First, the chain matrix$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$ of the circuit on the left side of the dotted line and the chain matrix $\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$ of the circuit on the right side should be calculated. Since these two circuits are in cascade, the composite network has the chain matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \text{X} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$$

$$= \begin{pmatrix} a_1 a_2 + b_1 c_2 & a_1 b_2 + b_1 d_2 \\ c_1 a_2 + d_1 c_2 & c_1 b_2 + d_1 d_2 \end{pmatrix}$$

The necessary and sufficient condition for the composite network to be symmetrical is that a = d or $a_1 a_2 + b_1 c_2 = c_1 b_2 + d_1 d_2$.

The expressions for a_1, b_1, c_1, d_1, a_2, b_2, c_2 and d_2 are then written in terms of λ, and the coefficients of corresponding terms are equated. Then, three simultaneous equations with three unknowns, i.e., C_5', C_7', and t, will be obtained and their values can be easily determined.

VII. CALCULATION OF AN EQUIVALENT LATTICE SECTION

Once C_5', C_7' and t have been determined, the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ of the composite network of Figure 7 is known. The characteristic function $K(\lambda)$ of the network may be computed according to the well known formula[2]

$$K(\lambda) = \frac{1}{2}(c + d - a - b).$$

Since $a = d$, we have

$$K(\lambda) = \frac{1}{2}(c-b).$$

Now the problem is reduced to the calculation of a simple 6th order crystal-capacitor lattice filter with the given characteristic function $K(\lambda)$. Since the method of that calculation is shown in most network synthesis textbooks, it is not described here.[2]

VIII. CALCULATION OF RESIDUAL ELEMENTS

With reference to Figure 6, after C_5', C_7' and t have been determined, the residual elements C_5'', C_7'' and t' can be calculated by the steps shown in Figure 8. Since the reactance transformations required in the process of Figure 8 are quite simple, the formulas are not shown here.

XI. NEGATIVE RESIDUAL ELEMENTS

In calculating the residual elements, it is assumed that $C_5 > C_5'$ and $C_7' > C_7$, otherwise C_5'' and C_7'' would be negative. In practice when the ladder networks are transformed section by section into lattices, negative residual elements appear at the last section. In order to make the filter realizable, C_1 and C_3 in Figure 7 must be changed into C_1' and C_3' such that $C_1 > C_1'$ and $C_3 < C_3'$. If C_1' and C_3' are chosen properly, the residual elements will become positive. In practice C_1' and C_3' may be chosen to make C_5'' and C_7'' negligible in the circuit, i.e., to make C_5'' very small and C_7'' very large, compared to the load.

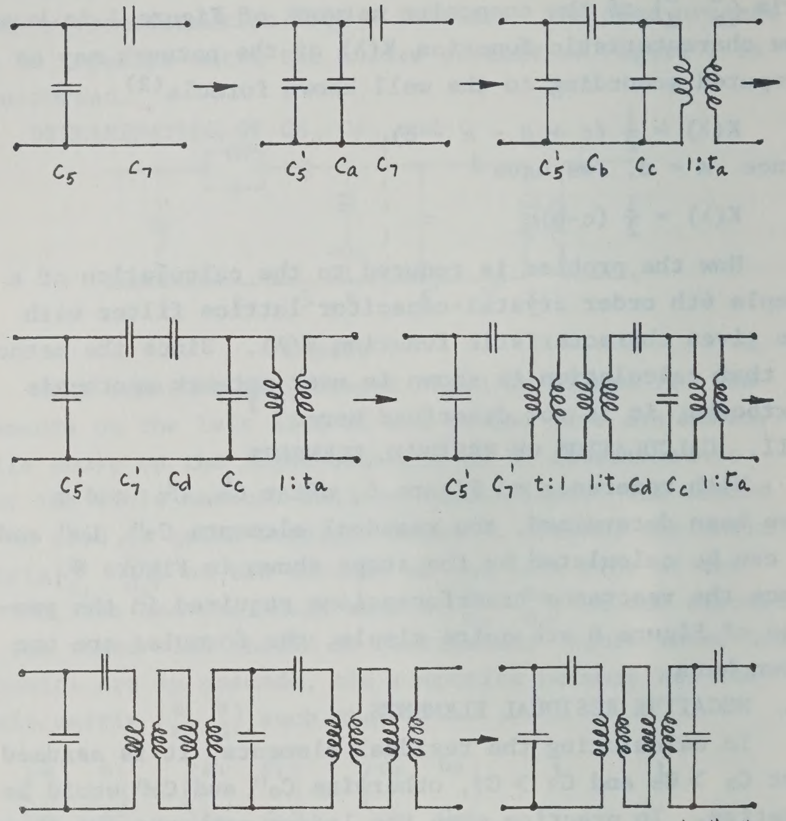

C_5 C_7
C_5' C_a C_7
C_5' C_b C_c $1:t_a$
C_5' C_7' C_d C_c $1:t_a$
C_5' C_7' $t:1$ $1:t$ C_d C_c $1:t_a$
C_5' C_7' $t:1$ $1:t$ C_e C_f $t_b:1$ $1:t_a$
C_5' C_7' $t:1$ $1:t'$ C_5'' C_7''

figure 8

With C_1' and C_3' known, a transformation similar to Figure 8 can be used to calculate the residual elements in front of the lattice section. This is possible because C_1' and C_3' instead of C_1 and C_3 have been used to make a lattice. As mentioned before, a symmetrical network cannot be formed by a cascade of non-identical symmetrical sub-networks, because some residual elements are always present either between lattice sections or at the end. It is these elements which make the whole filter symmetrical when they are combined with the lattice sections.

X. NUMERICAL EXAMPLE

As an illustration of the design procedures described above, a filter with the following specifications will be designed:

Center Frequency	10 Mc
Passband Width	17 Kc
Max. Passband Ripple	0.2 db
Stopband Width	29 Kc
Minimum Attenuation	33 db.

Since the response of the filter is very similar to the even order low pass filter, a quick check of response curve shows that C0441 will meet the requirements.

From the Glowatzki table,(3) all the Cauer parameters are obtained. After bandpass transformation, the locations of zeroes and poles can be determined as follows:

$\Omega_{01} = 1.0004000$

$\Omega_{02} = 0.9996000$

$\Omega_{03} = 1.000875$

$\Omega_{04} = 0.999125$

$\Omega_{\infty 1} = 1.0034375$

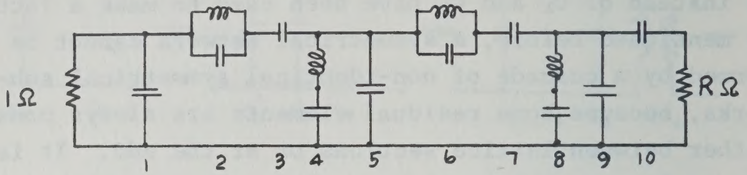
$\Omega_{\infty 2} = 0.9965625$

$\Omega_{\infty 3} = 1.0015625$

$\Omega_{\infty 4} = 0.9984375.$

With the locations of the zeroes and poles known, the constant factor h of $k(\lambda)$ is determined by the maximum ripple in the passband after $k = 1$ is assigned.

The characteristic function $K(\lambda)$ of the filter is then determined. It can be realized in ladder form as is shown in Figure 9.

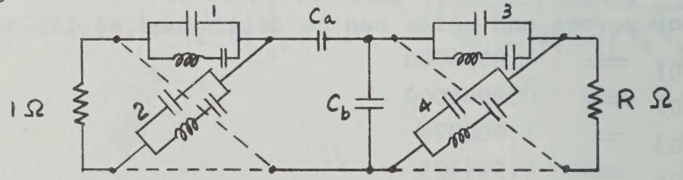


$C_1 = 0.6469195$
$C_2 = 46.308797$
$L_2 = 0.02144647$
$C_3 = 0.3738075$
$L_4 = 22.38074$
$C_4 = 0.0449901$
$C_5 = 6.395272$

$C_6 = 220.7614$
$L_6 = 0.004515654$
$C_7 = 0.54242277$
$L_8 = 239.80827$
$C_8 = 0.0041830602$
$C_9 = 0.96814182$
$C_{10} = 4.23030806$
$R = 0.9159$

Figure 9

From Figure 9, it is clear that this circuit cannot be realized by quartz crystals because of the low Cp/Cs ratio of the first three crystals required by the circuit. By using the method previously described, the network can be transformed into cascaded lattices as shown in Figure 10 below after all transformers are eliminated.



$L_{s1} = 389.2005$
$C_{s1} = 0.00257656$
$C_{p1} = 0.9123863$
$F_{s1} = 0.99860323$

$L_{s2} = 324.5075$
$C_{s2} = 0.00308166$
$C_{p2} = 1.0929384$
$F_{s2} = 0.99998885$

$L_{s3} = 157.5606$ $\quad L_{s4} = 564.6819$
$C_{s3} = 0.006348727$ $\quad C_{s4} = 0.001775625$
$C_{p3} = 2.13617812$ $\quad C_{p4} = 0.59582701$
$F_{s3} = 0.99984516$ $\quad F_{s4} = 0.99867105$
$C_a = 2.9480134$
$C_b = 0.398952928$
$R = 0.8845$

Figure 10

It is apparent that the above circuit can be implemented using quartz crystals without difficulty.

The response curve of the above filter is shown in Figure 11.

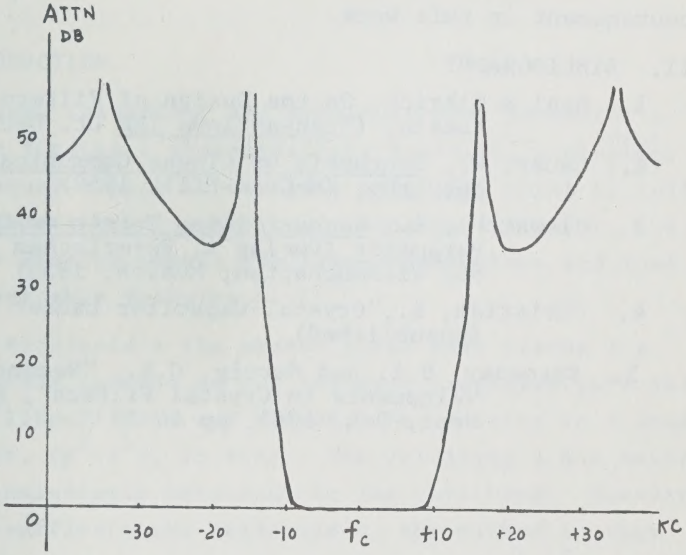


Figure 11

XI. CONCLUSIONS

The method described in this paper provides an approach for determining a realizable design of symmetrical filters with finite poles in a cascaded crystal-capacitor lattice form.

The limitation for practical realization of the filter depends on the C_p/C_s ratio of the crystals required. That ratio in turn depends on the relative bandwidth of the filter, and on the locations of the poles. Thus far, a filter with a 0.3% bandwidth, a 1:2.0 form factor, and an 80 db rejection in the stopband has been realized. When ceramic resonators become commercially available, the design method will undoubtedly have wider application.

XII. ACKNOWLEDGEMENT

The author wishes to thank Messrs. David C. Pinkerton, William P. Bartley and Harry P. Thomas for their encouragement in this work.

XIII. BIBLIOGRAPHY

1. Saal & Ulbrich, On the Design of Filters by Synthesis, (Transactions IRE Ct. Dec.1958)
2. Cauer, W., Synthesis of Linear Communication Networks, (McGraw-Hill, 1958)
3. Glowatzki, E., Sechsstellige Tafeln de Cauer - Parameter (Verlag A. Bayerischen Akademie der Wissenchaften; Munich, 1955)
4. Christian, E.,"Crystal-Capacitor Ladder Filters" (unpublished)
5. Kosowsky, D.I. and Hurtig, C.R., "Recent Developments in Crystal Filters", *Frequency*, Sept.-Oct. 1963, pp 44-53

DETERMINATION OF CHARACTERISTIC FUNCTIONS FOR CRYSTAL-CAPACITOR BANDPASS LADDER FILTERS WITH NEARLY CHEBYSHEV PASS-BAND ATTENUATION

M. Sablatash
Northern Electric Company
Research and Development Laboratories
Ottawa, Ontario, Canada.

A. INTRODUCTION

Although Schoeffler and Macdonald have presented a method for finding a characteristic function, $\phi(p)=\frac{F(p)}{P(p)}$, which is equal-ripple in the pass band, and which is suitable for ladder crystal-capacitor band-pass filters, the method (as will be shown) has some restrictions and some other undesirable features.(1),(2)

Like Macdonald's the method given here yields a ϕ with arbitrary (except at $p=0$ and $p=\infty$) imaginary-axis poles and, like Schoeffler's and Macdonald's, results in a real-zero factor, $(p^2-\sigma^2)$, in $F(p)$. The resulting ϕ has nearly perfect equal-ripple behaviour in the pass-band. However, unlike Schoeffler's and Macdonald's, the method in this paper yields a ϕ with a prescribed factor, $(p^2-\sigma^2)$, in $F(p)$, and with none of the undesirable characteristics. One example of such a ϕ which appeared in a recent article by Szentirmai may be found by this method.(3)

The method given by Theorem IV of a paper by Watanabe, for finding a ϕ with the above factor in F, is not efficient except for wide-band filters.(4) If the pass band is wide, however, the crystal-capacitance ratios are usually physically unrealizable. The method presented in this paper does not produce the degradation in attenuation and, because it allows the use of an arbitrary (within large limits) prescribed real-zero factor, enables (as will be discussed and illustrated by examples) the improvement of crystal-capacitance ratios.

Although Watanabe's Theorem 5 yields an F(p) (of odd degree) with a real-zero, $(p \pm a)$, and the method of obtaining it results in no undesirable properties of $\phi(p)$, the final network contains a minimum of one coil.(5)

It will be shown that a ϕ with a factor, $(p-a)(p+b)$, is suitable for the realization of crystal-capacitor ladder band-pass filters, and a method for obtaining such a ϕ (with nearly equal-ripple pass-band behaviour) will be presented.

All the above ϕ's, including those found by Schoeffler and by Macdonald, are easily derived, as will be shown, by applying Watanabe's general approximation theory.(6)

The potential analogy is heuristically derived and used collaterally with the mathematics to provide a lucid physical visualization of Watanabe's theory, and its application to the approximation problems treated specifically in this paper.(7)

To determine $\phi(p)$ from $w(p)$, the ideal transmission function, equation (16) in Watanabe's paper may be used, although the author prefers a procedure due to Bennett.(8),(9),(10),(11) This procedure is justified below. The resulting $\phi(p)$ is equal-ripple in the pass band, and has real zeros and real poles or pole branch points of second order. These pole factors are all cancelled after the real-zero factors are shifted into coincidence with them. The resulting $\phi(p)$ is nearly perfectly Chebyshev in the pass band and has the required, specified real zeros.

The position(s) of the real zero(s) is (are) degree(s) of freedom which may be utilized to advantage in the realization of the network. The effects of its (their) position on realization, network configuration, and component values are discussed and illustrated by examples.

B. DERIVATION OF THE GENERAL CHEBYSHEV PASS-BAND APPROXIMATION THEORY BY MEANS OF THE POTENTIAL ANALOGY

To derive Watanabe's general theory in potential analog terms the simplest cases are considered first and these generalized.(12),(13)

In the potential analogy for the simplest image-parameter low-pass propagation function a conducting plate is placed along the p-plane imaginary axis between -j and j and a positive point charge, $2\gamma_\infty$, is placed at $p=\infty$ (Figure 1). The complex potential in the s plane, where $s=u+jv$ $=\sinh^{-1}p$, is $w_\infty=2\gamma_\infty s$, for $u>0$, so in the p plane

$$w_\infty = 2\gamma_\infty \sinh^{-1}p = 2\gamma_\infty \ln(p+\sqrt{p^2+1}), \tag{2.1}$$

and

$$\coth\left(\frac{w_\infty}{2\gamma_\infty}\right) = \frac{\sqrt{p^2+1}}{p}. \tag{2.2}$$

Next, if one places positive point charges, γ_ν, only at $p=\pm j\omega_\nu$, and a conducting plate with negative charge, $-2\gamma_\nu$, along $[-j,+j]$, one finds that in the right-half s plane the complex potential, $w_\nu(s)$, is given (after removal of the plate and replacement of left-half plane charges by charges of opposite sign) by an infinite sum of logarithmic terms, from which one gets (Figures 2, 3(a) and 3(b))

$$\exp\left(\frac{w_\nu(s)}{\gamma_\nu}\right)= \frac{\cosh(s+a)}{\cosh(s-a)}, \qquad u>0, \tag{2.3}$$

where $a = \sinh^{-1}\omega_\nu$. In the p plane, with $m_\nu=\tanh a$,

$$\exp\left(\frac{w_\nu(p)}{\gamma_\nu}\right)= \frac{\sqrt{1+p^2} + m_\nu p}{\sqrt{1+p^2} - m_\nu p}, \tag{2.4}$$

$$\coth\left(\frac{w_\nu(p)}{2\gamma_\nu}\right) = \frac{\sqrt{1+p^2}}{m_\nu p}, \tag{2.5}$$

or

$$w_\nu(p) = 2\gamma_\nu \sinh^{-1}\left(\frac{\sqrt{\omega_\nu^2-1}\ p}{\sqrt{p^2+\omega_\nu^2}}\right). \tag{2.6}$$

In general, if there are positive point charges, $\gamma_{-\nu}$, at $p=\pm\sigma_{-\nu}$, (on the real axes), γ_ν at $p=\pm j\omega_\nu$, and γ_i at $p=\pm p_i$ and $p=\pm p_i^*$, the general complex potential (image propagation function) in the low-pass case is

$$w(p) = 2\gamma_\infty \sinh^{-1}(p) + \sum_{\nu'=1}^{k} 2\gamma_{-\nu'} \sinh^{-1}\left(\frac{\sqrt{\sigma_{-\nu'}^2+1}\;p}{\sqrt{\sigma_{-\nu'}^2-p^2}}\right)$$

$$+ \sum_{\nu=1}^{h} 2\gamma_\nu \sinh^{-1}\left(\frac{\sqrt{\omega_\nu^2-1}\;p}{\sqrt{\omega_\nu^2+p^2}}\right) + \sum_{i=1}^{g} 2\gamma_i \left[\sinh^{-1}\left(\frac{\sqrt{p_i^2+1}\;p}{\sqrt{p_i^2-p^2}}\right)\right.$$

$$\left. + \sinh^{-1}\left(\frac{\sqrt{p_i^{*2}+1}\;p}{\sqrt{p_i^{*2}-p^2}}\right)\right] + W_o, \qquad (2.7)$$

where Wo is a complex constant. For comparison with Cauer's work it should be noted that Equations (2.2) and (2.5) are q functions, that q′ functions, and q functions with real and complex one points are easily worked out, and that Cauer's composition rules for q and q′ functions are applicable.(14)

The $2\gamma_\infty$, $\gamma_{-\nu'}$, γ_ν and γ_i are residues at the poles of $\psi(p)=dw(p)/dp$. After one uses Watanabe's bilinear transformations from the p plane to the low-pass λ plane and employs both positive and negative point charges, the analogy of the $w(\lambda)$ with Watanabe's low-pass generalized ideal transmission function becomes clear.(15)

Watanabe's Theorem 1 defines two possible w(p)'s, one (the nonpole w(p)) analogous to the complex potential when only conducting plates, with positive or negative charge, are present along the $j\omega$-axis; and the other to the complex potential when positive and negative point charges are also permitted.(16),(17) Since only positive or negative charges are possible, the w(p)'s are completely general. They may be regarded as mappings of potential problems to orthogonal potential coordinates (on a Riemann surface), as represented by Figures 4 and 5.

Quantization of charge along the zero potential contours yields the zeros of the generalized characteristic function with equal-ripple pass-band behaviour,

$$\phi(\lambda) = \varepsilon \cosh w(\lambda), \qquad (2.8)$$

whose zeros occur where $w(\lambda) = j(\nu + \frac{1}{2})\pi$ (ν an integer).

Figures 6 and 7 depict quantization for w with and without poles, respectively.

In general, residues of $\frac{dw(p)}{dp}$ are analogous to point charges. With these analogies on hand, Watanabe's theory is easily explained and visualized. After fixing the poles with negative residues (according to the methods for producing real zeros which are described below), and adding the magnitude of their (negative) attenuation to the specifications, the rest of the poles and their (positive) residues are determined by well-known template or digital computer methods or with the aid of a potential analog device.(18),(19),(20),(21) Thus w(p) is determined, and F(p) can be found using equations in Watanabe's paper, or Bennett's procedure, which is described next.

C. THE USE OF BENNETT'S PROCEDURE IN FINDING THE NUMERATOR OF $\phi(p)$

The known w(p), regarded as a complex potential, may be mapped to the z plane, where(22)

$$z = \sqrt{\frac{p^2+\eta_c^{\,2}}{p^2+\eta_{-c}^{\,2}}}, \tag{3.1}$$

with $\eta_c = \omega_p/\omega_o$, $\eta_{-c} = \omega_{-p}/\omega_o$ and $\omega_o = \sqrt{\omega_p\omega_{-p}}$.

The conducting plate along the pass-band regions of the p plane maps into the whole of the z-plane imaginary axis, and the point charges along the p-plane axes map onto the z-plane real axis. Figure 8(a) depicts this mapping of w(p) for the numerical example to be given later.

Upon removal of the conducting plate in the z plane and replacement, according to the method of images, of the point charges in, say, the left-half of the z plane by charges of opposite sign, the exponential potential in the right-half of the z plane remains unchanged, and is given by

$$e^{\psi(z)} = \prod_{\nu=1}^{N} \left(\frac{z+z_\nu}{z-z_\nu}\right), \tag{3.2}$$

where the z_ν's corresponding to positive point charges have positive real parts and the z_ν's corresponding to negative point charges have negative real parts (Figure 8(b)). Then the function

$$\phi(z) = \frac{e^{\psi(z)} + (-1)^N}{2} = \frac{\prod_{\nu=1}^{N} (z+z_\nu) + (-1)^N \prod_{\nu=1}^{N} (z-z_\nu)}{2 \prod_{\nu=1}^{N} (z-z_\nu)}$$

$$= \frac{\mathrm{Ev} \prod_{\nu=1}^{N} (z+z_\nu)}{\prod_{\nu=1}^{N} (z-z_\nu)} \qquad (3.3)$$

has an equal-ripple behaviour along the whole of the z axis. This may also be shown by writing, for z imaginary (z=jv),

$$\phi(jv) = \frac{\cos\theta + j\sin\theta\,(-1)^N}{2}, \qquad (3.4)$$

whose absolute value varies between 0 and +2.

One has the choice, at this point, of transforming $\phi(z)$ to the p plane and performing the remaining calculations there, or of doing them all, including realization, in the z plane. In this paper the Chebyshev $\phi(z)$ is transformed to the p plane.

D. THE APPLICATION OF THE APPROXIMATION THEORY TO DETERMINATION OF $\phi(p)$ WITH SPECIFIED REAL ZEROS

If $\phi(p) = F(p)/P(p)$ has one specified real-zero factor, $p^2-\sigma^2$, and the (even) F(p) is one greater in degree than P(p), P(p) has only jω-axis zeros and σ is within a certain range, $\phi(p)$ may be used for finding a band-pass ladder network containing only crystals and capacitors. Such a $\phi(p)$ with nearly equal-ripple pass-band behaviour is found by starting with a w(p) which is illustrated (corresponding to the numerical example given later), typically, by Figure 9.

First, in setting up w(p) one selects a σ and places a negative charge of -1 at the points $\pm\sigma$. That is, the residue of dw(p)/dp at $\pm\sigma$ is -1. The residues at p=0 and p=∞ must each be unity if the poles there are to be simple and those at the other jω-axis poles are integers, usually unity.

The contributions of the poles at $\pm\sigma$, $p=0$ and $p=\infty$ are fixed,(23) so there remains only the positioning of the infinite, non-zero $j\omega$-axis poles by template, digital computer or potential analog methods. With the above signs of the residues one must have

$$\alpha = \mathrm{Re}[w(p)] = 0$$

at two pairs of points, one on either side of $\pm\sigma$. The imaginary part, β, of $w(p)$ is an odd multiple of $\frac{\pi}{2}$ on the real p axis on either side of $\pm\sigma$, because the quarter-encirclement of a unit charge at $p=0$ or $p=\infty$ results in a flux change of $\frac{\pi}{2}$, while the half-encirclement around any of the others results in a flux change of $\pm\sigma$ (or an integral multiple thereof). This is merely an application of Watanabe's Lemma 1 in potential analogy terms. Thus, on either side of $p=\pm\sigma$, $w(p) = 0+jk\frac{\pi}{2}$, where k is an odd integer, and

$$\phi(p) = \varepsilon \cosh w(p) = 0.$$

Hence, there will be a factor $(p^2-\sigma_1^2)(p^2-\sigma_3^2)$ in the numerator of $\phi(p)$ and a factor $(p^2-\sigma^2)$ in its denominator. This $\phi(p)$ is perfectly equal-ripple in the pass bands. Because all poles except the real pole contribute positive attenuation, σ_1^2 and σ_3^2 are close to σ^2, so that the factor $(p^2-\sigma_1^2)(p^2-\sigma_2^2)$ may be replaced by $(p^2-\sigma^2)^2$, and one of these cancelled with the denominator factor, yielding a $\phi(p)$ with a factor $(p^2-\sigma^2)$ and nearly perfect equal-ripple pass-band behaviour.

Schoeffler and Macdonald obtain a $\phi(p)$ with equal-ripple pass-band behaviour by changing, in Bennett's procedure, the sign of the z_ν corresponding to $p=0$ or $p=\infty$, to be opposite to that of the z_ν corresponding to the other poles.(24),(25) This corresponds to making the residue of $\frac{dw(p)}{dp}$ at $p = 0$ or ∞ equal to -1 -- that is, the charge placed there is -1, as illustrated by Figure 10. The resulting real-zero factor is not arbitrary, and appreciable movement (which may be necessary) disturbs the attenuation. There is also a minimum of α between the negative charge and the adjacent (positive) charge on the $j\omega$-axis, for $w(p) = 0+jk\pi$ (k an integer) at some point between them.

Finally, Watanabe's Theorem 4, which yields a factor, $p^2-\sigma^2$, but results in undesirable $\phi(p)$'s, as mentioned in the Introduction, is illustrated by Figure 11.

To obtain a $\phi(p)$ with a factor, $(p+a)(p-b)$, one begins by placing charges of $-\frac{1}{2}$ at two distinct pairs, $\pm\sigma_2$ and $\pm\sigma_3$, of real-axis points, and of 1 each at $p=0$ and ∞, as illustrated by Figure 12. Then $\alpha=0$ at $\pm\sigma_1$ and $\pm\sigma_4$, where $0>\sigma_1<\sigma_2<\sigma_3<\sigma_4<\infty$, and $\beta=k\frac{\pi}{2}$ (k an integer) there. Thus, there is a factor, $(p^2-\sigma_1^2)(p^2-\sigma_4^2)$, in the numerator of ϕ. Letting $a=\sigma_1^2/\sigma_2$ and $b=\sigma_4^2/\sigma_3$, cancelling the $\sqrt{(p^2-\sigma_2^2)(p^2-\sigma_3^2)}$ and noting that, for $p=j\omega$, $|\sqrt{p^2-a^2}|=|p\pm a|$, one obtains the desired ϕ with factor $(p-a)(p+b)$. As shown in the next section, factors $(p+a)(p+b)$ and $(p-a)(p-b)$ are not permissible.

Although there may be other cases, the author has not yet found others for band-pass filters.

The author has also used the above ideas for obtaining real-zero factors in ϕ's for double pass-band filters.(26)

E. REALIZABILITY AND REALIZATION

From $\phi(p)$ the numerator, $E(p)$, of the operating voltage ratio is obtained by well-known methods.(27),(28) From $E(p)$ and $F(p)$ one obtains the open- and short-circuit reactances from which the network is realized.(29),(30)

If there is a factor $(p^2-\sigma^2)$ in the numerator of ϕ, the open- and short-circuit reactances are

$$X_{10} = X_{20} = \frac{E_o}{E_e \pm F_e}$$

and

$$X_{1S} = X_{2S} = \frac{E_e \mp F_e}{Eo}.$$

After realization simple network transformations result in the desired crystal-capacitor ladder network.

As the examples of the next section illustrate increasing $|\sigma|$ causes the following:

1. Shift of zeros and poles of reactances towards higher frequencies. A great enough increase causes intermingling of transmission zeros with these zeros and poles. This does not preclude realizability, for this depends on the order of removal of transmission zeros. Thus, by removing a number of transmission zeros in the lower stop band first, one may shift the zeros and poles of the reactances below the transmission zeros in the upper stop band, and the latter may be realized, subsequently, in the usual way. One can examine, at each step, the locations of the zeros and poles of reactance in order to determine the next removal. One possibility is the realization of a transmission zero in the upper stop band as a shunt resonant circuit.
2. A decrease in the size of all capacitors.
3. A large increase in the crystal capacitance ratios (shunt divided by series capacitor) in series arms and a slight decrease of ratios in shunt arms. The reason for the decrease of the capacitors involved in partial and final removals of the transmission zeros at $p=0$ and $p=\infty$ follows from the shift of reactance zeros and poles towards infinity. Thus, partial removals of the pole of admittance at infinity are smaller and those of poles of impedance at zero are larger. The decrease of the other capacitors may be qualitatively explained by considering the change in the slope of the reactances.

No rigorous analysis of the effect of the position of the real zero is made here. Perhaps root locus analysis could be used for this purpose.(31)

If the factor, $(p-a)(p+b)$, is in the numerator of ϕ, the open- and short-circuit reactances are

$$X_{10} = \frac{Eo \mp Fo}{E_e \pm Fe}, \quad X_{20} = \frac{Eo \pm Fo}{E_e \pm F_e},$$

$$X_{1S} = \frac{E_e \mp Fe}{Eo \pm Fo} \text{ and } X_{2S} = \frac{Ee \mp Fe}{Eo \mp Fo}.$$

The same types of cancellation occur as before. No investigation has yet been made of the effect of $|a-b|$ upon realizability and element values.

Factors $(p+a)(p+b)$ and $(p-a)(p-b)$ are impossible for crystal ladder networks because the resulting X_{10} and X_{20} would have a zero at $p=0$ or a pole at $p=\infty$.

F. NUMERICAL EXAMPLES

Ladder networks for which ϕ has a factor $(p^2-\sigma^2)$ were synthesized to meet the following specifications:

$f_{-p} = 104.57$ kc $f_p = 107.795$ kc $A_p = 0.3$ db $Ro = 135.\ \Omega$

Pole frequencies: $f_1 = 102.892$ kc $f_2 = 104.00$ kc
$f_3 = 108.55$ kc $f_4 = 109.37$ kc.

The frequencies were normalized to $f_o = \sqrt{f_{-p} f_p}$, so that $\omega_o=1$. Bennett's procedure, as explained above, was included in a larger program. The σ were taken as $(\frac{30000.}{fo})$, $(\frac{90000.}{fo})$, and $(\frac{150000.}{fo})$ and three sets of networks synthesized. The numerical calculations, up to the point where the zeros and poles of the reactances are found, are shown only for the case in which $\sigma = (\frac{150000.}{f_o})$, although the zeros and poles of all three reactances are shown for comparison, as well as all three sets of networks.

For $\sigma = \frac{150000.}{f_o}$ and perfectly Chebyshev behaviour

$\phi=F(p)/P(p)=$

$$\frac{\begin{array}{l} p^{12}+0.0062499071098076485515641324p^{10} \\ -5.991459300075447596979775 4p^{8} \\ -4.021299042039817366967784p^{6} \\ +8.938105527416182528321248 4p^{4} \\ +11.943023736763107922172805p^{2}+3.9813298607611764620461 06 \end{array}}{\begin{array}{l} 0.2721418114280811695602561 6p^{11} \\ +0.5468311328878151747811594p^{9} \\ -0.54018782433704044264013282p^{7} \\ -2.1756103829408539073461701p^{5} \\ -1.904023782328478215850121 1p^{3}+0.5432961099454024744641423 5 \end{array}}$$

The squared roots of F(p) and $\frac{P(p)}{p}$ in the order found by Newton iteration, are:

Squared roots of F(p)	Squared roots of P(p)/p
-0.9715329765465l	-0.93928585895752
-0.98557368870180	-0.95962431701682
-1.0291728008181	-1.0454282256750
2.0484106887011	-1.0612824686072
-1.0138165273833	1.9962599050368
1.9454353976389	

The positive squared root of P(p) was dropped, and the two positive squared roots of F(p) replaced by this factor. Then the final $\phi(p)=F(p)/P(p)=$

$$\frac{\begin{aligned}&p^{10}+2.0038360884130p^{8}-1.9859734823747p^{6}\\&-7.9778498208714p^{4}-6.9824357712481p^{2}-1.9943960171327\end{aligned}}{\begin{aligned}&0.27210130231243p^{9}+1.0899346553667p^{7}\\&+1.6356854358412p^{5}+1.0899667159388p^{3}+0.27211649011119p\end{aligned}}.$$

The roots of F+P were found and those with positive real parts negated to obtain E(p)=

$$\begin{aligned}&p^{10}+2.878990991464803589036714 4p^{9}\\&+6.111111093520696995468781 6p^{8}\\&+11.5586640498608047989724 97p^{7}\\&+14.3295836873191222476099 03p^{6}\\&+17.3966879016200669550279 2p^{5}\\&+16.3201211525162344155168 25p^{4}\\&+11.6333105121406310563146 28p^{3}\\&+9.0960440331972521859942317p^{2}\\&+2.9162975127738409972524529p\\&+1.99439601713279353417662 96.\end{aligned}$$

The reactances $X_{10} = X_{20} = \frac{Eo}{E_e + F_e}$ and $X_{1S}=X_{2S}=\frac{Ee - Fe}{Eo}$ were formed next.

$$X_{10}=\frac{\begin{array}{l}2.878990991464803589036714457p^{8}\\+11.558664049860804798972497 1p^{6}\\+17.396687901620066955027919 93p^{4}\\+11.633310512140631058314628 99p^{2}\\+2.916297512773840997252452995\end{array}}{\begin{array}{l}2.00p^{8}+8.1149471819337086009690871p^{7}\\+12.343610204944325587198254 7p^{5}\\+8.3422713316448120348677026 2p^{3}\\+2.1136082619490884688698066 8p\end{array}}.$$

The squared roots of $\frac{Eo}{p}$ and $\frac{Ee + Fe}{p^2}$ for $\sigma = \frac{150000.}{fo}$, $\sigma = \frac{90000.}{fo}$ and $\sigma = \frac{30000.}{fo}$ are:

	Squared roots of $\frac{Eo}{p}$	Squared roots of $\frac{Ee + Fe}{p^2}$
$\sigma=\frac{150000.}{f_o}$	-1.04307899778	-1.04340297211
	-1.02409624587	-1.04257741600
	-0.982073339001	-1.00043641815
	-0.965583031652	-0.971056784688
$\sigma=\frac{90000.}{f_o}$	-1.03691766960	-1.03822010213
	-1.01904799015	-1.03323238260
	-0.977090491454	-0.993260648618
	-0.961494557882	-0.970082634020
$\sigma=\frac{30000.}{f_o}$	-1.03290745131	-1.03473920130
	-1.00839103241	-1.02386273980
	-0.971612610632	-0.984992185656
	-0.928044216813	-0.968691353522

The shifts in the zeros and poles and their positions relative to the squared transmission zeros, which are at -0.939285858, -0.9596243162, -1.0454282244 8 and -1.061282467284808 6, are to be noted.

Finally, ladders were realized using all possible orders of transmission zero removals. Two of these are shown, for each of the above real-zero positions, in Figures 13-15. The changes in element values and capacitance ratios are to be noted.

G. CONCLUSIONS

Specific applications of Watanabe's general theory of approximation have yielded characteristic functions with nearly perfect equal-ripple pass-band behaviour, specified poles, and the necessary two (specified) real zeros needed for realization by crystal-capacitor ladder band-pass filters. One of the solutions, yielding a factor, $(p^2-\sigma^2)$, of the characteristic function, was shown to be an improvement over those previously known which yield such a factor. The other yielded a factor $(p-a)(p+b)$. Both solutions are new in the literature.

A derivation of the approximation theory by means of the potential analogy was provided to aid in visualization of the functions, and was found collaterally useful in presenting the mathematical steps.

Bennett's procedure for finding the numerator of the characteristic function with Chebyshev pass-band behaviour was shown to be applicable to the cases in which $\frac{dw(p)}{dp}$ has positive and negative residues.

Finally, the effect of shifting the real zero on the reactance zeros and poles and network element values was observed in examples in which $\phi(p)$ had a factor $(p^2-\sigma^2)$, and it was found that increasing $|\sigma|$: (1) shifted the poles and zeros towards infinity, (2) decreased all the capacitors and increased all the inductances and (3) increased the capacitance ratios in series arms, but (4) decreased slightly the crystal capacitance ratios in the shunt arms. Only numerical results were presented. The effect of σ on the realizability and its possible use in realizing transmission zeros in the upper stop band by shunt arms, or those in the lower stop band by series arms, was noted. An analysis should be made of the above effects of the real-zero position. When the factor $(p-a)(p+b)$ is present, the effects of $|a-b|$ should be investigated.

ACKNOWLEDGEMENTS

The author is indebted to Mr. B.R. Smith and Mr. M. Stolarik for preparing the computer programs and working out the examples of Section VI, to Mrs. Shirley Hutter for typing the manuscript, and to Otfried Riml for critizing the original manuscript. He is also indebted to the National Research Council of Canada and to Professors E. Bridges and R.A. Johnson of the University of Manitoba, through whom he was able to lay the foundations for the approximation theory described herein.

REFERENCES

1. James D. Schoeffler, "A Solution to the Approximation and Realization Problems for Crystal Ladder Filters," 1964 IEEE International Convention Record, Part 1 (March 23-26, 1964), pp. 282-288.

2. John D. Macdonald, "The Insertion-Loss Design of Narrow-Band Crystal Filters," Paper delivered at a Seminar on Network Synthesis (The University of Ottawa, Jan. 23, 1964).

3. G. Szentirmai, "On the Realization of Crystal Band-Pass Filters," IEEE Transactions on Circuit Theory, Vol. CT-11, No. 2 (June, 1964), pp. 299-301.

4. Hitoshi Watanabe, "Approximation Theory for Filter-Networks," IRE Transactions on Circuit Theory, Vol. CT-9, No. 3 (September, 1961), pp. 341-356.

5. Ibid.

6. Ibid.

7. M. Sablatash, "The Development of a Semi-Automatic Potential Analog Apparatus, a General and Integrated Approximation Theory for Filters, and the Implementation of the Apparatus in its Application," A Report to the National Research Council of Canada, on work done under NRC Grant No. A738, to Prof. E. Bridges of the University of Manitoba, Winnipeg, Canada (August, 1963). Part of this work has been accepted as the author's M.Sc. thesis.

8. Watanabe, op. cit., p. 344.

9. Byron J. Bennett, Linear Phase Electric Filters, Technical Report No. 43, Electronics Research Laboratory, Stanford University (February 14, 1952).

10. Byron J. Bennett, "Synthesis of Electric Filters with Arbitrary Phase Characteristics," *1953 IRE National Convention Record*, Part 5, Circuit Theory, pp. 19-26.

11. R. Saal and E. Ulbrich, "On the Design of Filters by Synthesis," *IRE Transactions on Circuit Theory*, Vol. CT-5, No. 4 (December, 1958), pp. 297 and 309-313.

12. Watanabe, *op*. *cit*., pp. 341-356.

13. Sablatash, *op*. *cit*., Chapters V and VI.

14. Cauer, *op*. *cit*., pp. 228-291 and pp. 548-553.

15. Watanabe, *loc*. *cit*.

16. J.F. Klinkhamer, "Empirical Determination of Wave-Filter Transfer Functions with Specified Properties," *Phillips Research Reports*, Vol. 3 (1948), pp. 60-80 and pp. 378-400.

17. Watanabe, *op*. *cit*., p. 342. 18. *Ibid*., pp. 344-345.

19. E. Rumpelt, "Template Method for the Design of Electrical Wave Filters," *Teleg*. *u*. *Fernsprechtech*., Vol. 31 (1942), pp. 203-210.

20. B.R. Smith and G.C. Temes, "An Iterative Approximation Procedure for Automatic Filter Synthesis," *1964 IEEE International Convention Record*, Part 1 (March 23-26, 1964).

21. Sablatash, *loc*. *cit*.

22. Saal and Ulbrich, *op*. *cit*., p. 297 and pp. 309-313.

23. Watanabe, *op*. *cit*., pp. 344-345.

24. Schoeffler, *loc*. *cit*. 25. Macdonald, *loc*. *cit*.

26. Watanabe, *op*. *cit*., p. 348

27. Saal and Ulbrich, *op*. *cit*., pp. 286-287 and pp. 304-306.

28. Wilhelm Cauer. *Synthesis of Linear Communication Networks*. Toronto: McGraw-Hill Book Co. Inc., 1958, pp. 426-445.

29. Saal and Ulbrich, *op*. *cit*., p. 287.

30. Cauer, *op*. *cit*., pp. 421-446.

31. James D. Schoeffler, "On the Existence and Realization of Crystal Ladder Filters," *Proceedings of the First Allerton Conference on Circuit and System Theory* (University of Illinois, Nov. 15-17, 1963), pp. 259-274.

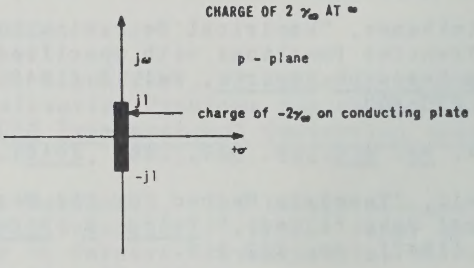

CHARGE OF 2 γ∞ AT ∞
jω
p - plane
j1
charge of -2γ∞ on conducting plate
+σ
-j1

FIGURE 1

POTENTIAL ANALOGY FOR SIMPLEST LOW-PASS
IMAGE-PARAMETER PROPAGATION FUNCTION

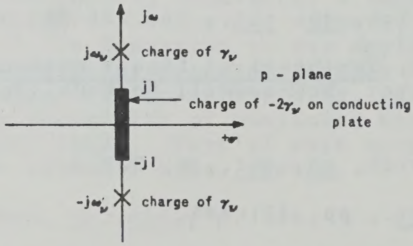

jω
jω_ν
charge of γ_ν
p - plane
j1
charge of -2γ_ν on conducting plate
+σ
-j1
-jω'_ν
charge of γ_ν

FIGURE 2

POTENTIAL ANALOGY FOR LOW-PASS IMAGE-PARAMETER
PROPAGATION FUNCTION, IN p - PLANE

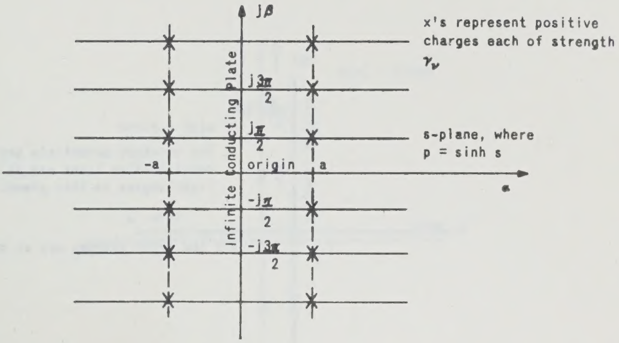

jβ
x's represent positive charges each of strength γ_ν
Infinite Conducting Plate
$j\frac{3\pi}{2}$
$j\frac{\pi}{2}$
s-plane, where p = sinh s
-a
origin
a
α
$-j\frac{\pi}{2}$
$-j\frac{3\pi}{2}$

FIGURE 3 (a)

MAPPING TO s-PLANE OF POTENTIAL PROBLEM FOR TWO COMPLEX CONJUGATE CHARGES ON IMAGINARY p-AXIS AND CONDUCTOR ALONG PASS BAND

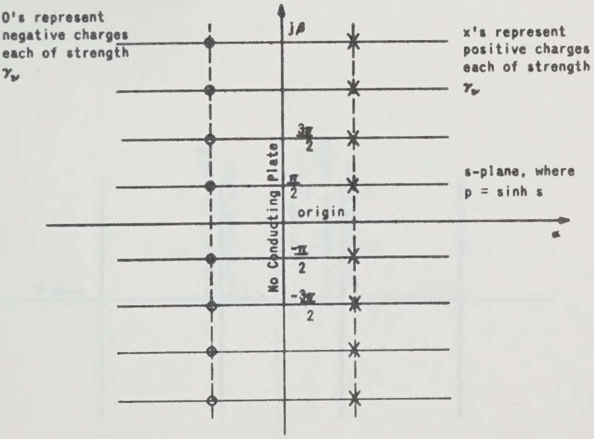

O's represent negative charges each of strength γ_ν
jβ
x's represent positive charges each of strength γ_ν
$\frac{3\pi}{2}$
$\frac{\pi}{2}$
No Conducting Plate
s-plane, where p = sinh s
origin
α
$-\frac{\pi}{2}$
$-\frac{3\pi}{2}$

FIGURE 3 (b)

POTENTIAL PROBLEM EQUIVALENT TO THAT OF FIGURE 3 (a) IN RIGHT HALF PLANE

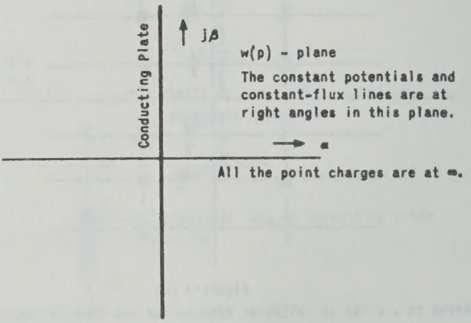

Conducting Plate
jβ
w(p) - plane
The constant potentials and constant-flux lines are at right angles in this plane.
α
All the point charges are at ∞.

FIGURE 4

W-PLANE REPRESENTING THE COMPLEX POTENTIAL CO-ORDINATES CORRESPONDING TO w(p) WITH (LOGARITHMIC) POLES

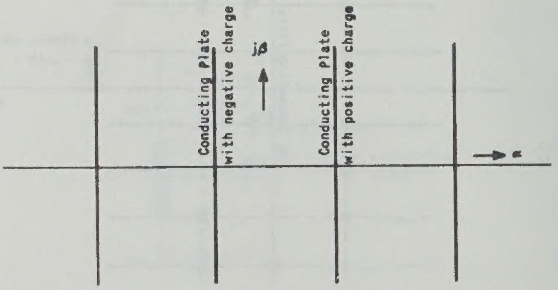

Conducting Plate
with negative charge
jβ
Conducting Plate
with positive charge
α

FIGURE 5

W-PLANE REPRESENTING THE COMPLEX POTENTIAL CO-ORDINATES (AND THEIR IMAGES, WHICH GIVE RISE TO A RIEMANN SURGACE) CORRESPONDING TO NON-POLE w(p). THE MAPPING DEFINES AN ELLIPTIC SINE OR TANGENT FUNCTION IN THE LOW-PASS CASE.

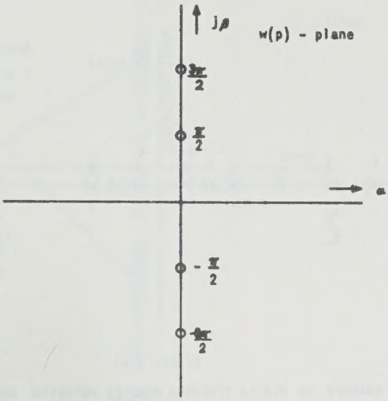

jβ
w(p) - plane
3π/2
π/2
α
- π/2
-3π/2

FIGURE 6

QUANTIZATION IN w-PLANE CORRESPONDING TO FIGURE 4 TO OBTAIN $\phi = \cosh w$

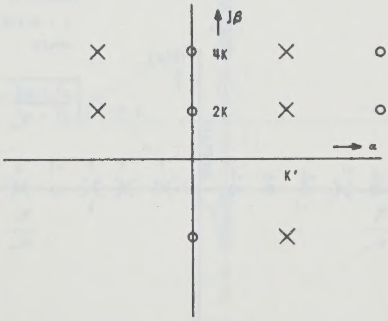


FIGURE 7

QUANTIZATION IN w-PLANE CORRESPONDING TO FIGURE 5 TO OBTAIN $\phi = j\ t_n\ [w,L]$

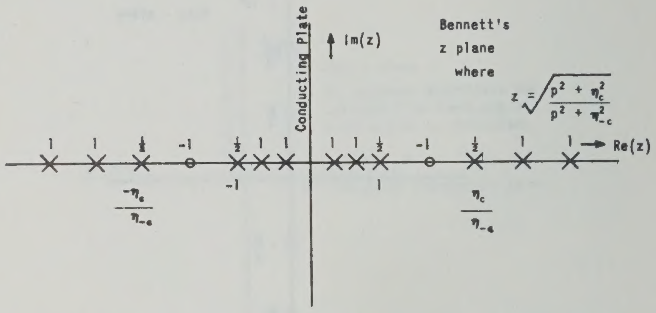

Conducting Plate
Im(z)
Bennett's
z plane
where
$z = \sqrt{\frac{p^2 + \eta_c^2}{p^2 + \eta_{-c}^2}}$
Re(z)
1
1
½
-1
½
1
1
1
1
½
-1
½
1
1
-1
1
$-\eta_c$
η_{-c}
η_c
η_{-c}

FIGURE 8(a)

SKETCH OF CHARGES AND PLATES YIELDING COMPLEX POTENTIAL ANALOGOUS TO IDEAL TRANSMISSION FUNCTION FOR NUMERICAL EXAMPLE ON VISIBLE SHEETS OF RIEMANN SURFACE IN BENNETT'S z PLANE

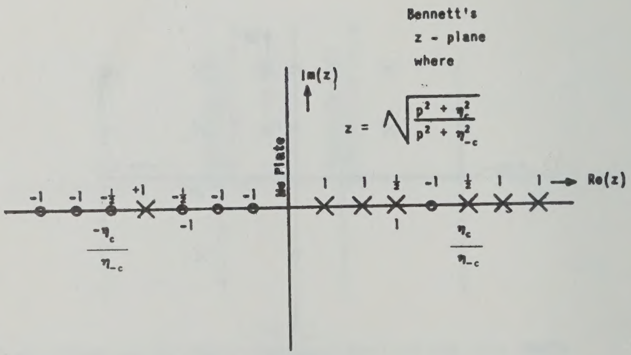

Bennett's
z - plane
where
Im(z)
$z = \sqrt{\frac{p^2 + \eta_c^2}{p^2 + \eta_{-c}^2}}$
No Plate
Re(z)
-1
-1
-½
+1
-½
-1
-1
1
1
½
-1
½
1
1
-1
1
$-\eta_c$
η_{-c}
η_c
η_{-c}

FIGURE 8(b)

SKETCH OF CHARGES (NO PLATES) YIELDING THE SAME COMPLEX POTENTIAL IN RIGHT HALF PLANE AS IN FIGURE 8(a)

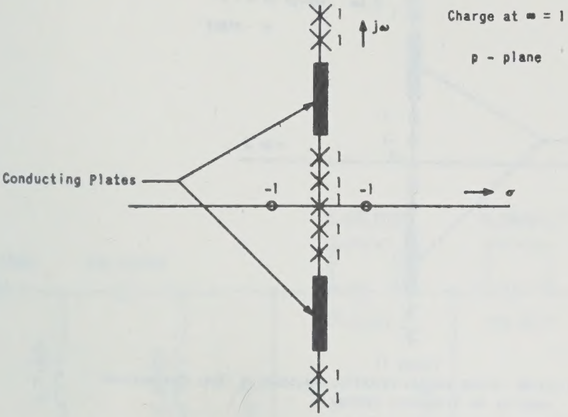

Charge at ∞ = 1
p - plane
jω
Conducting Plates
-1
-1
σ
1
1
1
1
1
1
1
1
1

FIGURE 9
SKETCH OF CHARGES AND PLATES CORRESPONDING TO FIGURES 9(a) and 9(b)
POINT CHARGES ON VISIBLE RIEMANN SURFACE ONLY

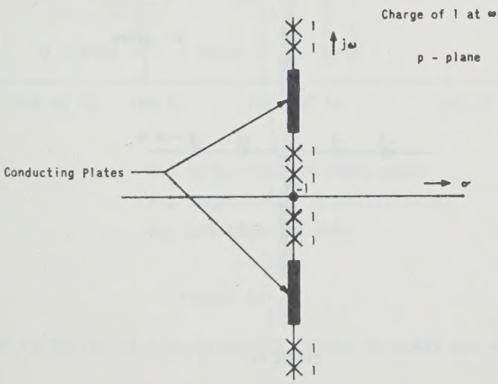

Charge of 1 at ∞
p - plane
jω
Conducting Plates
-1
σ
1
1
1
1
1
1
1
1

FIGURE 10
SKETCH OF CHARGES AND PLATES YIELDING TYPICAL COMPLEX POTENTIAL ANALOGOUS TO IDEAL TRANSMISSION FUNCTION FOR SCHOEFFLER'S METHOD

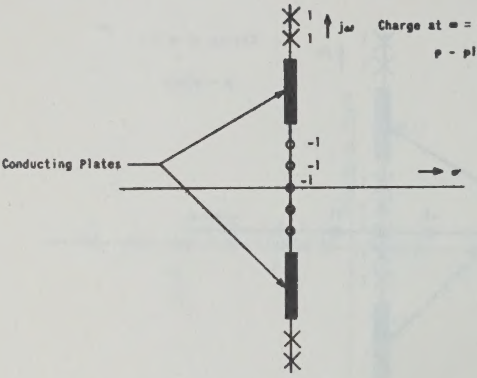

1
1
jω
Charge at ∞ = 2
p - plane
Conducting Plates
-1
-1
-1
σ

FIGURE 11
CHARGES AND PLATES YIELDING TYPICAL COMPLEX POTENTIAL ANALOGOUS TO IDEAL TRANSMISSION FUNCTION FOR WATANABE'S THEOREM 4

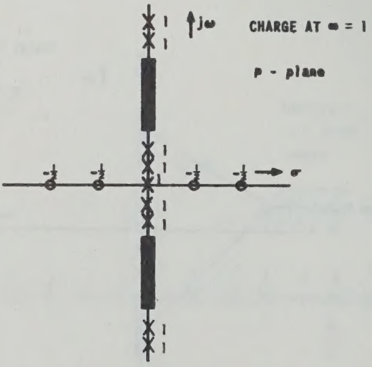
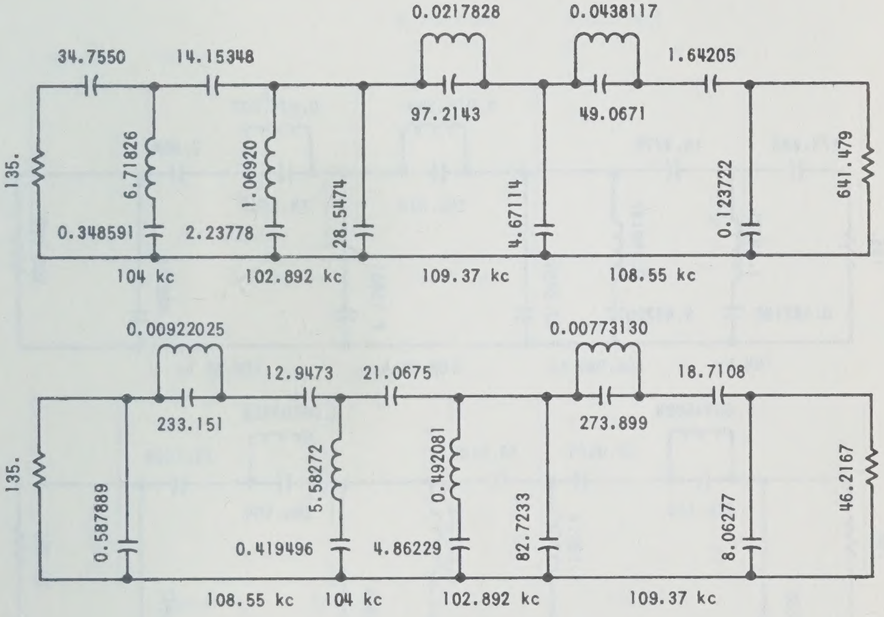

1
1
jω
CHARGE AT ∞ = 1
p - plane
1
1
$-\frac{1}{2}$
$-\frac{1}{2}$
1
$-\frac{1}{2}$
$-\frac{1}{2}$
σ
1
1
1
1

FIGURE 12

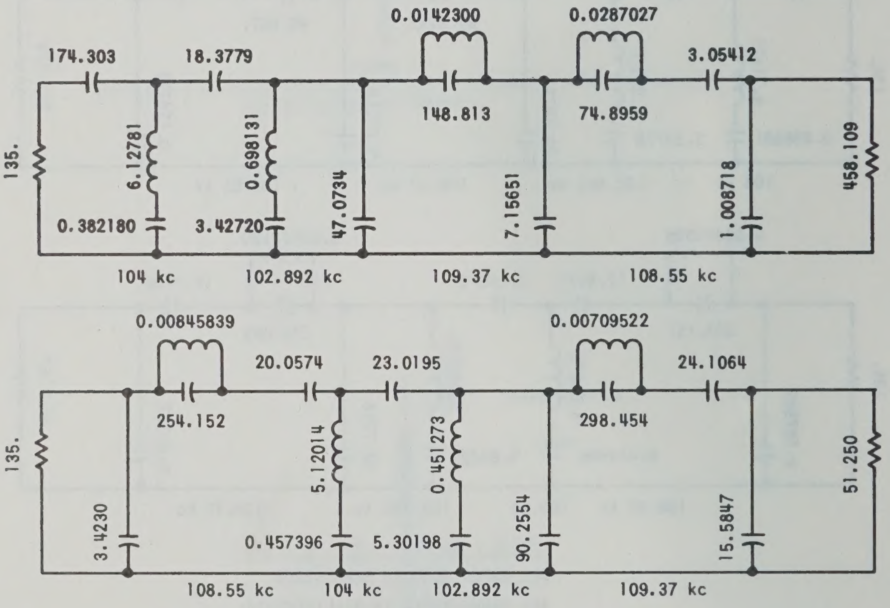
CHARGES AND PLATES YIELDING TYPICAL COMPLEX POTENTIAL ANALOGOUS TO IDEAL TRANSMISSION FUNCTION INITIALLY USED IN FINDING ϕ (p) WITH FACTOR (p-a)(p+b)


0.0217828
0.0438117
34.7550
14.15348
1.64205
97.2143
49.0671
135.
6.71826
1.06920
28.5474
4.67114
0.123722
641.479
0.348591
2.23778
104 kc
102.892 kc
109.37 kc
108.55 kc
0.00922025
0.00773130
12.9473
21.0675
18.7108
233.151
273.899
135.
0.587889
5.58272
0.492081
82.7233
8.06277
46.2167
0.419496
4.86229
108.55 kc
104 kc
102.892 kc
109.37 kc

ALL CAPACITORS IN NANOFARADS
ALL INDUCTANCES IN MILLIHENRIES
ALL RESISTORS IN OHMS

FIGURE 13

REALIZATIONS OF POTENTIAL CRYSTAL-CAPACITOR LADDER NETWORKS FOR $\sigma = \frac{150000.}{f_o}$


0.0142300
0.0287027
174.303
18.3779
3.05412
148.813
74.8959
135.
6.12781
0.698131
47.0734
7.15651
1.008718
458.109
0.382180
3.42720
104 kc
102.892 kc
109.37 kc
108.55 kc
0.00845839
0.00709522
20.0574
23.0195
24.1064
254.152
298.454
135.
5.12014
0.451273
3.4230
0.457396
5.30198
90.2554
15.5847
51.250
108.55 kc
104 kc
102.892 kc
109.37 kc

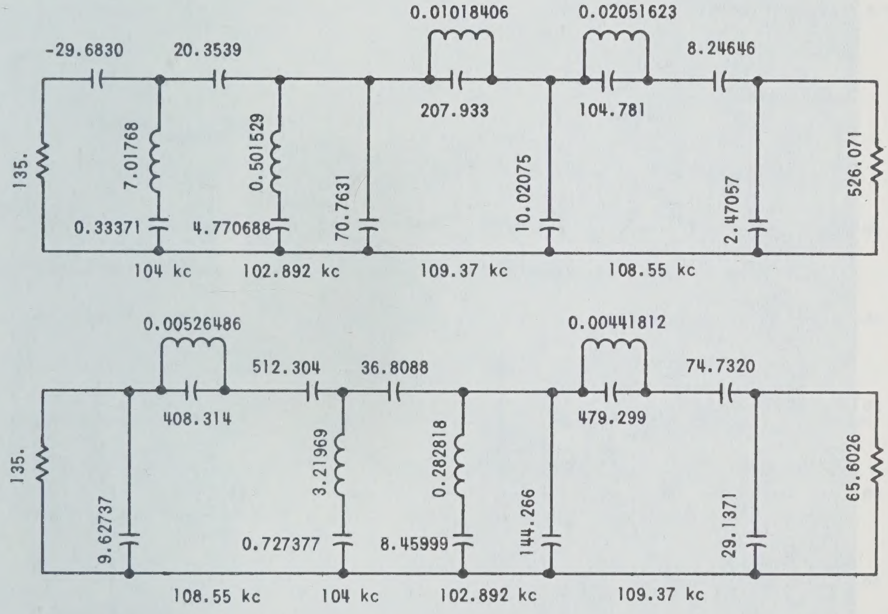
ALL CAPACITORS IN NANOFARADS

ALL INDUCTANCES IN MILLIHENRIES

ALL RESISTORS IN OHMS

FIGURE 14

REALIZATIONS OF POTENTIAL CRYSTAL-CAPACITOR LADDER NETWORKS FOR $\sigma = \frac{90000.}{f_o}$


0.01018406
0.02051623
-29.6830
20.3539
8.24646
207.933
104.781
135.
7.01768
0.501529
70.7631
10.02075
2.47057
526.071
0.33371
4.770688
104 kc
102.892 kc
109.37 kc
108.55 kc
0.00526486
0.00441812
512.304
36.8088
74.7320
408.314
479.299
135.
3.21969
0.282818
9.62737
144.266
29.1371
65.6026
0.727377
8.45999
108.55 kc
104 kc
102.892 kc
109.37 kc

ALL CAPACITORS IN NANOFARADS
ALL INDUCTORS IN MILLIHENRIES
ALL RESISTORS IN OHMS

FIGURE 15

REALIZATIONS OF POTENTIAL CRYSTAL-CAPACITOR LADDER NETWORKS FOR $\sigma = \frac{30000.}{f_o}$

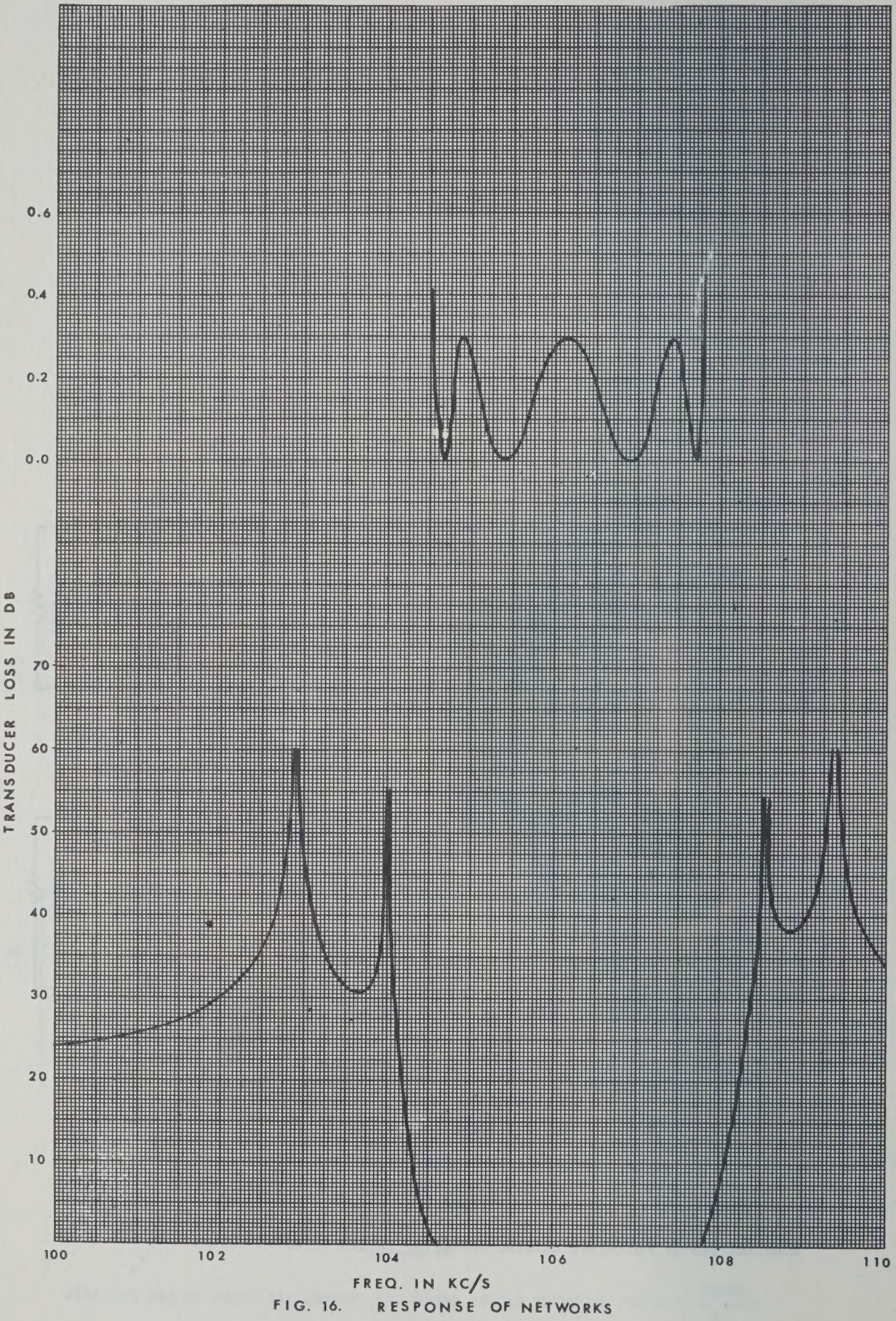


FIG. 16. RESPONSE OF NETWORKS

COMPUTER SOLUTION OF THE NETWORK REALIZATION PROBLEM*

D. A. Calahan
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

I. INTRODUCTION

When a network design (synthesis) problem arises which clearly requires a computer solution, the usual procedure, time and money allowing, is to write a program specifically for that network or class of networks. Although the extravagance is necessary in some design cases to avoid numerical accuracy problems and/or to avoid unacceptable computation times resulting from a more general program, nevertheless the number of "smaller problems" requiring no such sophistication is quite large. Also, a large number of problems, particularly in active networks, are usually not attacked with computer methods in part because the topologies vary widely, making it uneconomical to write a special computer program.

One may therefore inquire:

1) what should be the nature of a general design program.

2) what are the numerical approaches and problems involved.

---

* This work was supported by AFOSR Grant AF 177-63.

3) what are the network size limitations of such a program.

A set of answers to the above three questions is the subject of this paper.

II. THE APPROACH

Typically, the network designer has chosen a network configuration which, he feels, should yield a desired frequency response but which fails because of loss or other parasitics. He also has an estimation (not necessarily good) of the element values. Thus, he requires an alignment of the element values, a job suited to the computer.

It is certainly possible to accept any desired frequency response data as input to the program and then by repeated adjustment of elemental values and by matrix inversion of a complex-valued node admittance matrix, hopefully converge on the desired response. An alternative approach is to assume a bit more knowledge on the part of the designer -- namely, that he knows the actual network function T(s) desired (or perhaps just the dominant frequencies). Then, one may attempt to match coefficients of T(s).

The latter technique has been chosen here, because:

1) It is intuitive that one should in a general program profit from separating the realization problem from the approximation problem; for the frequency response is a natural function of the poles and zero locations rather than the coefficients of T(s), while the coefficients are linear functions of any particular element value. Indeed, experience of the author and others[1] has indicated the former methods which accept frequency response data tend to converge to local minima unless one uses special convergence

techniques inconsistent with the generality of the program.

2) The requirement that the designer specify T(s) is possibly no more stringent than that he exercise care in choosing the frequency data points in the former method.

III. NUMERICAL TECHNIQUES

The basic program is divided into two parts:

1) assembly of the coefficients of T(s) in terms of the elements

2) iteratively changing the element values

It is clear that topological methods are naturally suited to implementing (1). Although it is not clear that the most efficient methods of finding the trees of a graph have been determined, one may adopt any of several well-known methods.

The Newton-Raphson method[2] is employed to iteratively select the required element value changes that result in the desired coefficients. That is, given coefficients of a polynomial and desired coefficients a_i^d $(i = 1,\cdot\cdot n)$, error functions are defined,

$$E_i = K a_i^d - a_i \qquad i = 1, n \tag{1}$$

where K a constant independent of i. Then the desired roots are obtained if and only if $E_i = 0$. If there are precisely n - 1 variable elements e_i, then the elemental changes are given by

$$\sum_{j=1}^{n-1} \frac{\partial E_i}{\partial e_j} \triangle e_j = - E_i \qquad i = 1, n \tag{2}$$

or

$$- \sum_{j=1}^{n-1} \frac{\partial a_i}{\partial e_j} \triangle e_j + a_i^d \triangle K = a_i - K a_i^d \tag{3}$$

or

$$A \, \Delta e = B, \tag{4}$$

The Δe_j's are then formed by matrix inversion. Note that since the a_i's are multi-linear in the network elements, the partial derivatives are found simply from

$$\frac{\partial a_i}{\partial e_j} = \frac{\text{tree admittance products of } a_i \text{ containing } e_j}{e_j}, \tag{5}$$

Hence, the entries in A are known to the same precision as the coefficients*; also, once the latter are known, then obviously

$$\frac{\text{time to calculate A}}{\text{time to calculate coefficients}} \approx \frac{\text{time for single division } (40\ \mu\text{ sec})}{\text{time for single multiplication } (25\ \mu\text{ sec})}. \tag{6}$$

Thus, the networks size limitations are approximately the same as those of the topological network analysis program on which the method is based, since the latter is usually limited by the tree-finding procedure.

Typically, it is desired to specify certain equality side constraints, such as requiring an element to converge to a prescribed value, or requiring ratios of elements to be fixed (such as the R/L of an inductor). For each such constraint, an additional element must be made variable.

APPLICATION TO DESIGN OF LOSSY LADDERS

Consider the even order doubly terminated, ladder network of Figure 1, where loss is to be inserted (R_i, G_i) while R_o and R_t are held constant. Consider further

---

*which in turn are known to the same precision as the element values for RLC grounded networks.

1) the semi-uniformly lossy case ($G_i = 0$). There are a total of n poles to be held invarient as loss is inserted, or n + 1 coefficients to be kept proportional. In contrast, there are 3n/2 variable elements and a single undetermined constant of proportionality associated with the coefficients. Thus, a total of

$$\frac{3n}{2} + 1 - (n + 1) = \frac{n}{2} \qquad (7)$$

side constraints are possible. Forcing all inductors equally lossy yields

$$R_i - (d)(L_i) = 0 \qquad i = 1, 3, \cdot \cdot n - 1 \qquad (8)$$

a total of n/2 constraints and precisely the number required.

2) the uniformly-lossy case ($R_i/L_i = G_i/C_i = d$). Equation 7 now becomes

$$(2n + 1) - (n + 1) = n \qquad (9)$$

requiring n side constraints. To those of Equation 8 we add

$$G_i - d \; C_i = 0 \qquad i = 2, 4, \cdot \cdot n \qquad (10)$$

yielding a total of n side constraints.

It may be concluded that the classical problem of design of lossy ladders fits the algorithm proposed for any even order n (a similar statement applies for n odd). Such a problem will be termed to have a feasible solution, although an actual solution is of course not guaranteed.*

The above is an example of a general class of networks where insertion of an element does not increase the number of poles (n) of a given network function. In such a case, if all (m) reactive elements are made variable, then the number of equations required to match poles only is n + 1 and the number of unknowns is m + 1. Hence, addition of another variable element (such as a resistance to represent loss) allows at least one more constraint to be imposed on the element values (since $m \geq n$). Thus, the problem of predistortion of bandpass LC networks also has a feasible solution.

---

*For example, A may be singular

THE "DOMINANT POLE" CASE

Although the parasitics contained in LC networks are resistive and hence usually fulfill the requirements noted above, those associated with active networks are usually reactive and nearly always create new poles and/or zeros. To be useful in these cases, the algorithm must be extended so that only prescribed (usually "dominant") roots are maintained. In such a case, one would prescribe

$$P(s) = \sum_{i=0}^{p} b_i s^i \tag{11}$$

and then allow

$$Q(s) = \sum_{i=0}^{q} x_i s^i \tag{12}$$

to be arbitrary.

Note that, if D(s) is the polynomial defined by

$$D(s) = P(s) \; Q(s) \tag{13}$$

$$= \sum_{i=0}^{m} a_i s^i \qquad m = p + q - 1 \tag{14}$$

then

$$a_i = \sum_{j=0}^{i} b_j x_{i-j} \qquad \begin{array}{l} i - j \leq q. \\ j \leq p \end{array} \tag{15}$$

Hence defining

$$ERR_i = -a_i + \sum_{j=0}^{i} b_j x_{i-j} \qquad i = 1, \ldots m \tag{16}$$

yields, with n variable elements, equations

$$- \sum_{j=1}^{n} \frac{\partial a_i}{\partial e_j} \triangle e_j + \sum_{j=0}^{i} b_j \triangle x_{i-j} = a_i - \sum_{j=o}^{i} b_j x_{i-j}$$

$$i = 1, \ldots m \qquad (17)$$

which require (to make A square)

$$n + q = m + \text{of constraints.} \qquad (18)$$

It is interesting to note that, as the number of arbitrary roots increases (q), the number of required variable elements decreases--a quite reasonable result!

BOUNDS ON ELEMENTS

Clearly, there is a certain minimum number of variable elements required to make A of Equation 4 square. If more elements are made variable, it is usually with the intent of forcing an element to a prescribed value, thus adding constraint equations and forcing A to be square again. However, for some cases only bounds on elements are prescribed so that an inequality constraint is imposed and A is not square. A particularly interesting example of this is the case when a positive element is to be grown between two nodes. Thus, to Equation 2 one would add $\triangle e_j \geq 0$. Another case occurs when choice of a particular set of variable elements results in convergence, but with negative element values. In such a case, the computer can be used to pick an initial set of variable elements by imposing the conditions

$$\triangle e_j \geq - K_1 e_j \qquad K_1 \leq 1 \qquad (19)$$

and attempting to minimize (using linear programming) the derivative of some weighted function of ERR_i. Assuming a feasible and finite solution, the $\triangle e_j$'s which satisfy the equality sign in Equation 19 are immediately included in the list of variable elements. The remaining elements are somewhat arbitrarily chosen until A is square and the iteration proceeds.

One may ask why the linear programming algorithm is abandoned after the initial selection of element values. There are two principal reasons:

1) The number of variable elements is usually desired as small as possible.

2) According to the theory of linear programming, some $\triangle e_j$'s would always satisfy Equation 19 with an equality sign, so the process would never converge unless K_1 were decreased in some predetermined manner.

MINIMIZATION OF ZERO SENSITIVITY

Given the polynomial

$$D(s) = \overset{n}{\underset{i=1}{\pi}} (s - z_i) \qquad (20)$$

the zero sensitivity to e_j is

$$\mathcal{S}^{z_i}_{e_j} \equiv \frac{\partial z_i}{\frac{\partial e_j}{e_j}} \qquad (21)$$

$$= e_j \frac{\partial D(s)}{\partial e_j} \div \frac{N(s)}{s - z_i} \Bigg|_{s = z_i} . \qquad (22)$$

But the partial derivatives of the coefficients of D (s) are contained in A of Equation 4, so that $\mathcal{S}^{z_i}_{e_j}$ is readily calculated. One may then define a sensitivity criterion for a given element e_j as

$$\mathcal{S}_j = \sum_{i=1}^{n} \left[(\mathrm{Re}\ \mathcal{S}_{e_j}^{z_i})^2 + (\mathrm{Im}\ \mathcal{S}_{e_j}^{z_i})^2 \right]. \tag{23}$$

Then, one could choose a desired value $\mathcal{S}_j^d$ for the sensitivity and again define an error criterion

$$\mathrm{ERR} = -\mathcal{S}_j + \mathcal{S}_j^d \tag{24}$$

so that an equation of the form

$$2\sum_j \left\{ \sum_{i=1} \left[\left(\mathrm{Re}\ \mathcal{S}_{e_j}^{z_i} \right) \left(\frac{\partial \mathrm{Re}\ \mathcal{S}_{e_j}^{z_i}}{\partial e_j} \right) + \left(\mathrm{Im}\ \mathcal{S}_{e_j}^{z_i} \right) \left(\frac{\partial \mathrm{Im}\ \mathcal{S}_{e_j}^{z_i}}{\partial e_j} \right) \right] \right\} \Delta e_j = -\mathcal{S}_j^d + \mathcal{S}_j \tag{25}$$

is developed to be added to those of Equation 3.

RELATION TO "CONTINUOUS EQUIVALENCE"

It has been previously shown[3,4] that the equivalence of a network with respect to specified input and output ports could be maintained by any solution of the equation

$$\frac{d\,[y]}{dx} = [a]\,[y] + [y]\,[b] \tag{26}$$

where [a], [b] are square matrixes with appropriate rows and columns of zeros, and [y] is the branch or nodal admittance matrix. An alternative procedure to generating equivalent networks [in fact, the same equations are developed using either method] is to form the network function of interest and then to differentiate each coefficient with respect to x, viz.,

$$\frac{d\,a_i}{dx} = K\,a_i^d \tag{27}$$

where K is a constant. It is clear that any solution of Equation 27 will yield coefficients proportional to the original. Expanding Equation 27 in terms of the elements then produces equations very similar to Equation 3 with differentials replacing differences.

Computationally, each method [Equation 3 or 27] requires inversion of a matrix for solution. However, the integration step size necessary to achieve a given elemental accuracy requires an excessive number of matrix inversions, thus making the differential form much less desirable.

THE COMPUTER PROGRAM

The input data consists of a list of elements and values defining the topology of the network with a label to denote the variable elements. Elements which are offered as possibilities for "growing" are also entered, but with zero (or infinite) value. Next, if the response of the initial network is to be altered, the coefficient of the desired network function are entered [either numerator, denominator, or both]. Finally, any equality constraints on the elements are entered.

The output is a list of elemental values at every iteration, and the resultant coefficients, poles and zeros whenever convergence is obtained.

The flow chart of the program is given in Figure 2. Several points require elaboration:

1) When a parasitic element is present or when an element is to be "grown" between two nodes, it is labeled so that any new poles or zeros created by the element are considered "non-dominant." Thus, the dominant frequencies [which are the only ones retained invariant] are chosen implicitly by such labeling. It is also possible

to prescribe a like number of dominant frequencies in lieu of those of the original network.

2) When an excess number of elements are made variable, so that the computer must select an initial set or must attempt to eliminate negative elements then, assuming a feasible solution of the linear programming problem, the selection of the set proceeds in the following order:

a. elements in the constraint equation are chosen variables.

b. if an element $\triangle e_j$ satisfies Equation 19 with an equality sign, then e_i is chosen unless that element is denoted as a parasitic. Thus, when an initial set of elements is being chosen, elements being grown between nodes which could tend to grow with negative values are rejected, whereas normal elements, already positive but which tend to grow in a negative direction, are chosen. Thus, a chance is taken that these elements will eventually grow negative; however, the network function is typically very sensitive to such elements and keeping them a constant would usually place realizeability constraints impossible to satisfy by sucessive iteration.

c. elements are chosen in order of the size of the $\triangle e_i$'s which result from solution of the linear programming problem.

RESULTS*

It is only at this point that the term "realization program" will be justified. For it happens that the technique converges so well and from such poor initial guesses that it can be used in place of standard realization techniques. Thus, one can generate entire tables of filter designs similar to that in ref. [5] from a single low order design by "guessing" reasonable values for elements. Convergence then requires only several iterations. Similarly, one can generate tables using lossy elements.

The neutralization problem of bandpass transistor amplifiers [Figure 3] has also been solved, but taking into account transistor parasitics usually ignored in such designs. This required the use of the "dominant" pole option, since the parasitics created extraneous poles.

Employing the option to "grow" elements between nodes, new filter topologies have been obtained from existing ones [Figure 4]. Also, the classical problem obtaining "equal-element values" fits the algorithm, and some elliptic filters have been signed with equal inductors [but tolerating some loss] [Figure 5]. In the latter example, the program was offered the possibility of growing positive resistors between nodes 1-4, 2-4, and 3-4; the latter two were then routinely disregarded when they satisfied Equation 19 with the equality sign.

---

* An IBM 7090 Fortran program, together with an instruction manual, has been prepared, and is available from the author at nominal cost.

CONCLUSION

A scheme for the iterative design of networks of arbitrary configuration has been proposed, and the results of this procedure has been presented. Converence is so rapid and from such poor initial design that the term "realization technique" is felt to be justified. Even new topologies can be generated, leading possibly to new synthesis schemes.

If one criticism can be made of the procedure it is that the method is too precise; i.e., although convergence to an exact Tchebycheff or Butterworth characteristic may be impossible, an acceptable response may still be possible if only the corresponding poles were known. Accordingly, one way define "transitional" polynomials with coefficients that are linear functions of the standard polynomials. Then the constant(s) of proportionality become unknown(s) which can be determined by the Newton-Raphson iteration.

ACKNOWLEDGEMENT

The author would like to thank H. J. Orchard for his remarks concerning the extraordinary convergence properties of the coefficient matching technique. Also, the author acknowledges the use of the CDC 1604 computer of the Coordinated Science Laboratory, University of Illinois.

REFERENCES

1. Murata, T. "The Use of Adaptive Constrained Descent in System Design," IEEE Conv. Record, pt. 1, pp 296-306, 1964.

2. Hildebrand, F. B., Introduction to Numerical Analysis, McGraw-Hill, 1956.

3. Schoeffler, J. D., "The Synthesis of Minimum Sensitivity Networks," Trans. IEEE, vol. CT-11, pp 271-276, June 1964.

4. Calahan, D. A., "Computer Generation of Equivalent Networks," IEEE Conv. Record, pt. 1, pp. 2, pp. 330-337, 1964.

5. Weinberg, L., Network Analysis and Synthesis, McGraw-Hill, 1964.

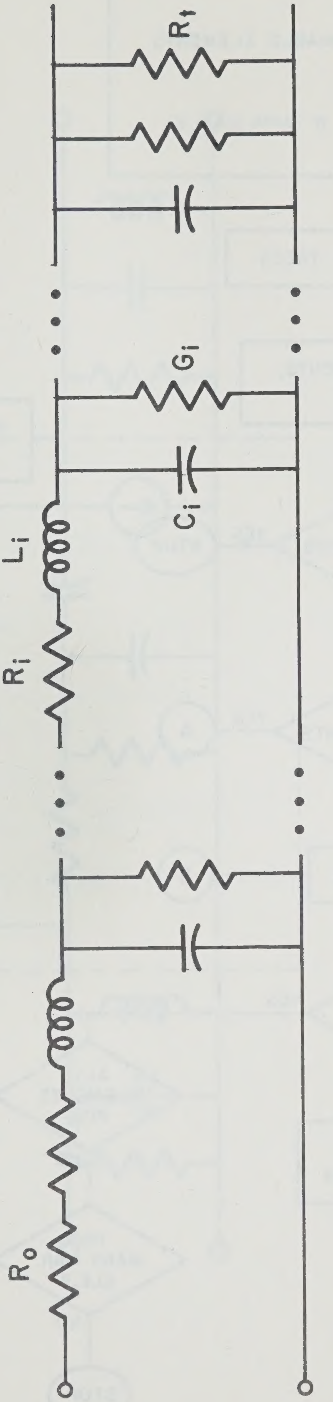

R_t
G_i
C_i
L_i
R_i
R_o

FIGURE I

Lossy Low-pass Ladder

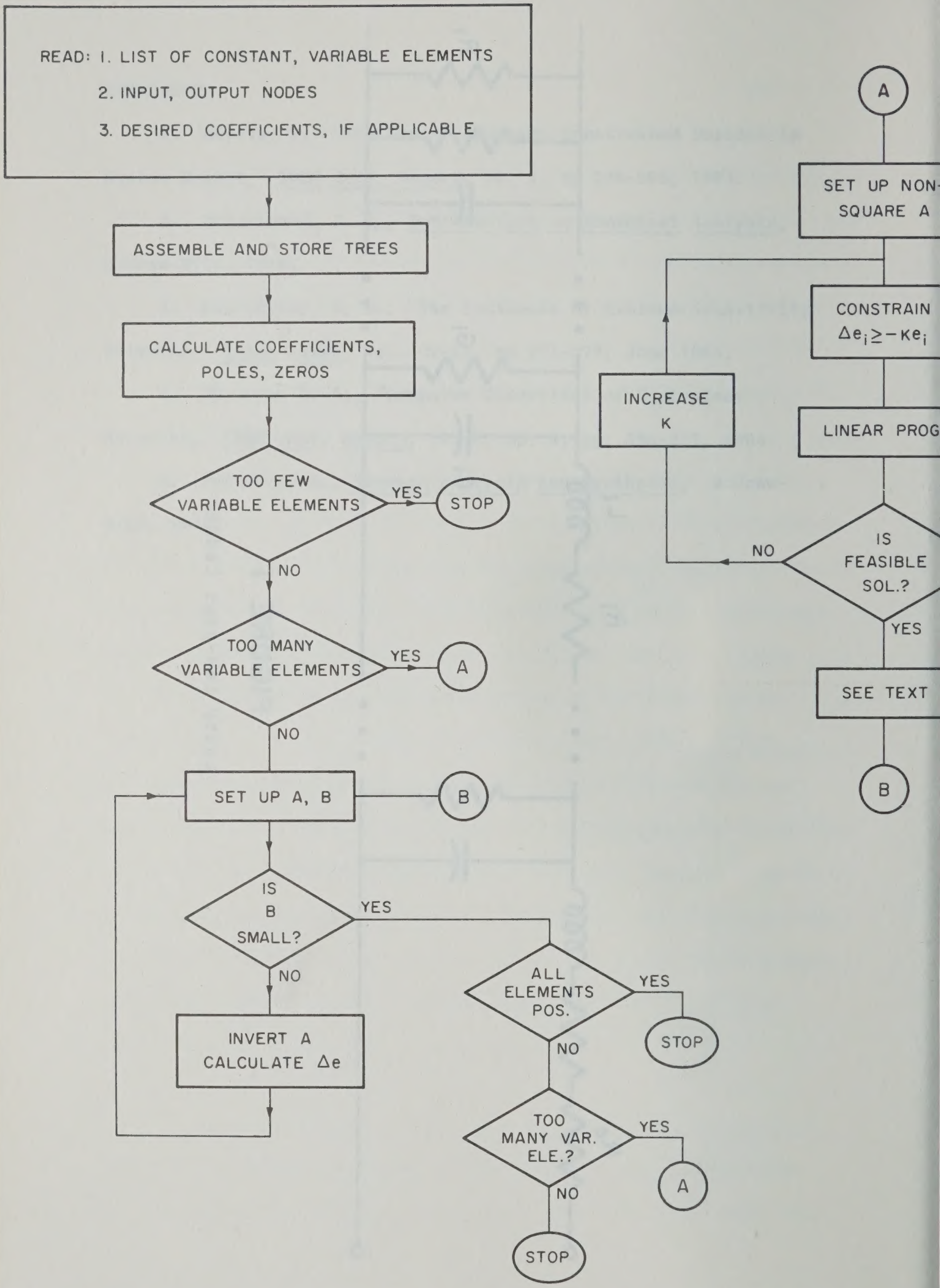

READ: 1. LIST OF CONSTANT, VARIABLE ELEMENTS
2. INPUT, OUTPUT NODES
3. DESIRED COEFFICIENTS, IF APPLICABLE
ASSEMBLE AND STORE TREES
CALCULATE COEFFICIENTS, POLES, ZEROS
TOO FEW VARIABLE ELEMENTS
YES
STOP
NO
TOO MANY VARIABLE ELEMENTS
YES
A
NO
SET UP A, B
B
IS B SMALL?
YES
NO
INVERT A CALCULATE Δe
ALL ELEMENTS POS.
YES
STOP
NO
TOO MANY VAR. ELE.?
YES
A
NO
STOP
A
SET UP NON-
SQUARE A
CONSTRAIN
$\Delta e_i \geq -\kappa e_i$
INCREASE K
LINEAR PROG
IS FEASIBLE SOL.?
NO
YES
SEE TEXT
B

FIGURE 2

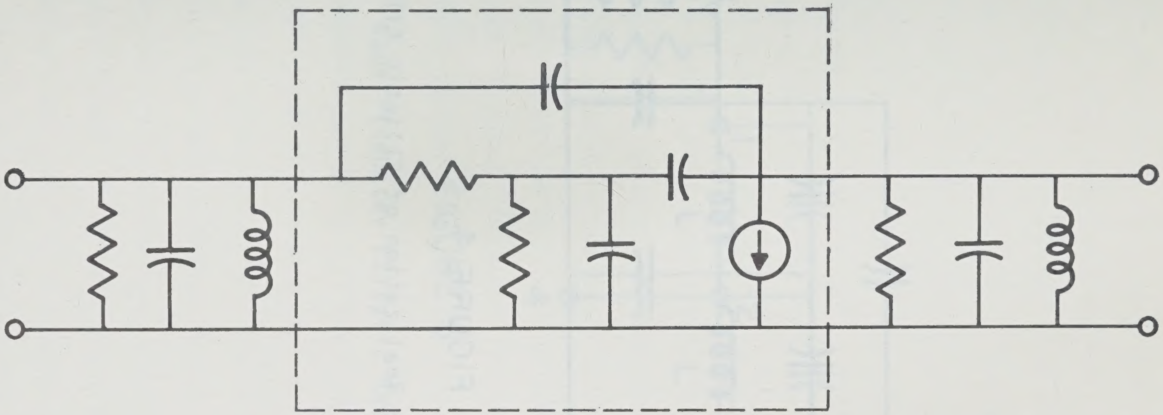


FIGURE 3

Bandpass Amplifier

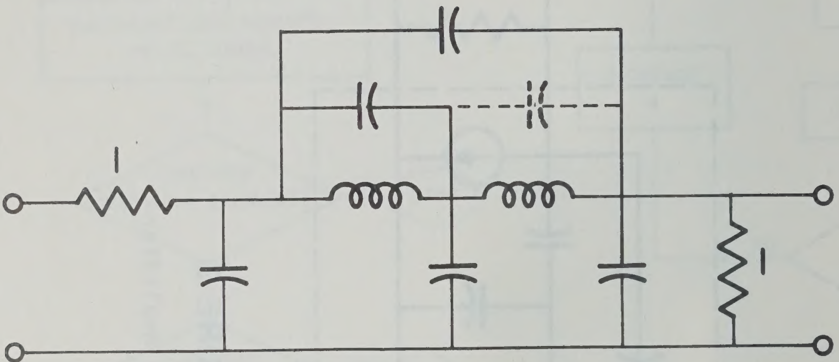


FIGURE 4

Alternative Realization of Elliptic Filter

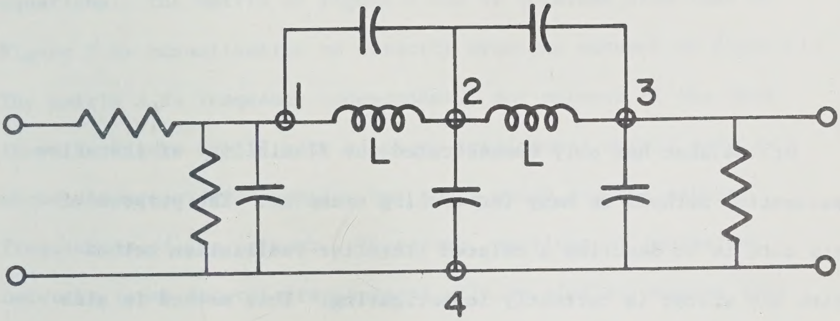


FIGURE 5

Equal-Inductor Elliptic Realization

DISCUSSION

Thomas G. Marshall, Jr.
Radio Corporation of America
Princeton, New Jersey

Dr. Calahan has ably demonstrated the flexibility of iterative realization methods in many interesting examples. The purpose of this note is to describe a related iterative realization method which the writer is currently investigating. This method is also related to those of Desoer,(1,2) Mitra,(2) Temes,(3) and Schoeffler,(4) but is sufficiently different that a description might prove of interest to the readers of this record. The written record was not available at this writing, so comparisons with Dr. Calahan's work will be necessarily limited.

Realization is accomplished in this method by similarity transformations. The input and output immittances are not invariant in this method, so a much broader class of transformations is possible than in Caur's method.(4) The transformation can be initiated from the response function as well as from an existing network, so complete realization is possible as well as equivalent network generation. An unusual aspect of this method is that polynomials are not formed. The objectives and techniques of this method can be illustrated by considering the problem of realizing the simple lowpass ladder network of Figure 1. It is assumed that the natural frequencies are prescribed and that, in addition, it is desired to specify the dissipation constants of the branches as freely as possible.

Bashkow(5) has shown that the matrix in Figure 2 can be formed by alternately writing Kirchhoff's voltage and current equations. The matrix of Figure 3 can be obtained from that of Figure 2 by normalization or directly from the network of Figure 1. The matrix A is frequency independent. Any network of the form shown in Figure 1 will have a unique tridiagonal matrix such as A whose characteristic equation has roots which are the natural frequencies of the network. Therefore, matrices of equivalent networks (same natural frequencies) will be similar because they have the same characteristic equations. The problem of generating equivalent ladder networks is therefore, in this formulation, one of generating similarity transformations upon tridiagonal matrices. A corollary property of these equivalent networks is that the sum of the dissipation constants, the trace of A, is constant.

A matrix similar to A may be written directly by placing negatives of the natural frequencies, denoted S_i, on the main diagonal as in Figure 4. Therefore, a similarity transformation exists which can transform the dissipation and coupling constants of a ladder network into the natural frequencies and vice versa. The lower diagonal of minus ones is included for convenience and not necessity.

One possibility for effecting the transformation from a given network or set of natural frequencies is to try to make the transformation in small steps. A method related to Dr. Calahan's has been used to do this. Assume that in a given transformation it is specified that the dissipation constants be given increments $\Delta\alpha_{ii}$, subject to the constraint that $\sum \Delta\alpha_{ii} = 0$. It is desired to determine the resulting n-1 $\Delta\alpha_{i,i+1}$. Therefore, only n-1 equations are required.

These equations are obtained by specifying response requirements (for example that there be no change) at particular frequencies. These frequencies are often chosen to be the natural frequencies, although other frequencies can also be chosen. The accuracy problems associated with polynominals are avoided by obtaining the n-1 equations in this manner rather than by specifying requirements upon polynominal coefficients. Polynominals are, in fact, not formed in this procedure.

The equation corresponding to the $j^{\underline{th}}$ frequency is

$$\sum_{i=1}^{n-1} \frac{\partial L}{\partial \alpha_{i,i+1}} \Delta\alpha_{i,i+1} = - \sum_{i=1}^{n} \frac{\partial L}{\partial \alpha_{ii}} \Delta\alpha_{ii} + \Delta L(j) \quad (1)$$

where $\Delta L(j)$ is any desired correction or change in the response at the $j^{\underline{th}}$ frequency. The presence of the $\Delta L(j)$ term permits corrections to be made for any errors introduced by previous transformations as well as permitting a response variation to be purposefully introduced. If j is a complex frequency, the real and imaginary parts of (1) yield two equations. If n-1 independent equations cannot be found in this manner, the right hand vector above cannot be arbitrarily specified, and it may not be possible to obtain the desired dissipation constants by this method.

The derivatives of the loss function (input over output) are calculated rather than those of the gain function, because they are, in general, simpler. This difference is particularly great in ladder networks where both the loss function and its derivatives are linear fuctions of the branch immittances. For a general network let the $i^{\underline{th}}$ branch immittance be I_i, and number the input branch 1 and the output branch n. Then

$$L = \frac{\Delta}{\Delta_{1n}} \quad \text{and}$$

$$\frac{\partial L}{\partial I_i} = \frac{\Delta_{1n}\Delta_{ii} - \Delta\Delta_{1n,ii}}{\Delta_{1n}^2} = \frac{\Delta_{1i}\Delta_{in}}{\Delta_{1n}^2} \qquad (2)$$

For ladder networks, $\Delta_{1n} = 1$ and Δ_{1i}, Δ_{in}, and Δ are respectively the continuants K (i + 1, n), K (1, i-1), and K (1,n).(6) Therefore,

$$\frac{\partial L}{\partial L_i} = \frac{\partial K(1,n)}{\partial I_i} = K(i+1,n)\ K(1,i-1) \qquad (3)$$

In a similar manner the partials of the off diagonal elements of the matrix A + sI in Figure 3 may be shown to be

$$\frac{\partial K(1,n)}{\partial \alpha_{i,i+1}} = K(i+1,n)\ K(1,i-1) \qquad (4)$$

It is seen that these expressions are not singular at the natural frequencies.

The K(i+1,n) for i = n, n-1, ..., 1 are easily determined by recursion (Muir, op. cit.). This is equivalent to the normal procedure of assuming the output of a ladder to be unity and calculating the branch variables successively to give the input. The K (1,i-1) may be determined in a similar manner corresponding to starting with branch 1. Hence, except for multiplying, all the partial derivatives are determined by just twice the effort of a normal loss calculation.

A Fortran program was written to test the method, and the program was tested on an RCA 301 computer. A 10th order doubly terminated ladder network was realized by transformation from the natural frequencies in 27 iterations, and several lower order networks were realized having specified positive and negative dissipation constants in all branches. 988 words of fixed storage were required for the object program and subroutines plus $n^2 + 9n$ words for data. About 1 minute per iteration was required for the 10th order network. Of this 6 seconds was required to solve the linear system and the remainder was divided about equally between determining the partial derivatives and printing out data. An RCA 301 computer was employed (add time is .33 milli-sec.).

It is thought that the method described herein is a useful iterative realization method for ladder networks. It is hoped that some of the techniques will be useful for the more general circuits considered by Dr. Calahan.

The writer would like to thank Dr. Calahan for discussing his conference paper with him. He would also like to thank Messrs. F. M. Brock and H. F. L. Cameron, Broadcast & Communications Products Div., RCA for their encouragement during this study.

REFERENCES

1. C. A. Desoer, "Network Design by First-Order Predistortion Technique", IRE Trans. on Circuit Theory, Vol. CT-4 (September, 1957) pp. 167-170.

2. C. A. Desoer and S. K. Mitra, "Design of Lossy Ladder Filters by Digital Computer", IRE Trans. on Circuit Theory, Vol. CT-9 (September, 1961) pp. 192-201.
3. G. C. Temes, "First-Order Estimation and Correction of Parasitic Loss Effects in Ladder Filters", IRE Trans. on Circuit Theory, Vol. CT-9 (December, 1962) pp. 385-399.
4. J. D. Schoeffler, "The Synthesis of Minimum Sensitivity Networks", IRE Trans. on Circuit Theory, Vol. CT-11 (June 1964) pp. 271-276.
5. T. R. Bashkow, "A Note on Ladder Network Analysis", IRE Trans. on Circuit Theory (correspondence), Vol. CT-8 (June 1961) pp. 168-169.
6. T. Muir, rev. ed. by W. H. Metzler, A Treatise on the Theory of Determinants. Dover Publications, Inc.: New York, New York (1960) pp. 516-542.

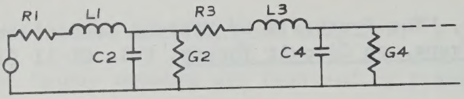


FIGURE 1

$$\begin{bmatrix} R_1 + sL_1 & 1 & & & \\ -1 & G_2 + sC_2 & 1 & & \\ & -1 & R_3 + sL_3 & 1 & \\ & & -1 & G_4 + sL_4 & 1 \\ & & & -1 & \ddots \end{bmatrix}$$

FIGURE 2

$$\begin{bmatrix} \frac{R_1}{L_1} + s & \frac{1}{L_1 C_2} & & \\ -1 & \frac{G_2}{C_2} + s & \frac{1}{C_2 L_3} & \\ & -1 & \frac{R_3}{L_3} + s & \frac{1}{L_3 C_4} \\ & & -1 & \ddots \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & & \\ -1 & a_{22} & a_{23} & \\ & -1 & a_{33} & a_{34} \\ & & & \ddots \end{bmatrix} + sI$$

$$= A + sI$$

FIGURE 3

$$\begin{bmatrix} -s_1 & & & & \\ -1 & -s_1^* & & & \\ & -1 & -s_2 & & \\ & & -1 & -s_2^* & \\ & & & & \ddots \end{bmatrix}$$

FIGURE 4

ON THE OPTIMUM LEAST-SQUARES APPROXIMATION TO THE IDEAL LOWPASS*

C. Pottle and J. S. Thorp
School of Electrical Engineering
Cornell University
Ithaca, N. Y.

I. INTRODUCTION

There now exist a large number of approximations to the so-called "ideal lowpass" - an unrealizable filter which passes with linear phase those frequencies below a cutoff frequency while completely rejecting those above cutoff. Most of these approximations yield realizable filters whose magnitude or phase characteristic approximates that of the ideal lowpass in an optimum fashion according to some chosen criterion. This class of filters does not yet include those filters which approximate the ideal lowpass in the least integral-squared-error sense. That the parameters of least-squares filters have not yet been found would appear odd, since this criterion has been employed almost to the exclusion of all others in control and communication system theory, where its chief attraction has been the mathematical tractability it offers. The explanation seems to be that as yet no analytical

---

* This research was supported by the Ford Foundation through a grant to the Cornell University College of Engineering.

technique exists for locating the optimum pole positions of this class of filter, and one must resort to numerical techniques. Although this paper does not provide such analytical methods, it presents some useful numerical methods, and it is hoped that the results obtained will provide a clue to the problem of discovering analytical techniques.

Using the least integral-squared-error criterion, the parameters of a new class of filters approximating the ideal lowpass will be presented. Since for this particular criterion the index of performance is the same in both the time and frequency domain, an equivalent title for the paper might be "On the Optimum Realizable Least-Squares Approximation to the Delayed Sin t/t Function."

The results obtained here by numerical methods seem to lead to several conjectures which once made may be not too difficult to prove.

II. THE PROBLEM

The ideal lowpass filter has the frequency characteristic shown in Figure 1.

$$F(j\omega) = \begin{cases} e^{-j\omega t_o} , & |\omega| \leq 1 \\ 0 , & |\omega| > 1 \end{cases} \qquad (1)$$

We wish to approximate this characteristic by a filter with the transfer function

$$H(s) = K \frac{\prod_{k=1}^{M} (s - z_k)}{\prod_{k=1}^{N} (s - s_k)} \qquad (2)$$

where $H(s)$ is a rational function whose numerator degree is at least one less than the degree of the denominator. The number N of poles to be found in the approximation, and that their real parts be negative, is specified at the outset. The parameters to be chosen in the approximation are the zero and pole positions. The quantity t_o (the group delay) can be specified in advance or included as a parameter. The approximation must be done in such a way as to minimize the index of performance

$$I_1 \overset{\Delta}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty}\Big[F(j\omega) - H(j\omega)\Big]\Big[F(j\omega) - H(j\omega)\Big]^{*} d\omega \quad . \qquad (3)$$

In the time domain, the approximation would seek to minimize the index

$$I_2 \overset{\Delta}{=} \int_{-\infty}^{\infty}\Big[f(t) - h(t)\Big]^2 dt \quad . \qquad (4)$$

By Parseval's Theorem, when the parameters are chosen so as to minimize I_1, I_2 is also minimized and is equal to I_1. The time function to be approximated is given by

$$f(t) = \frac{1}{\pi}\,\frac{\sin\,(t - t_o)}{t - t_o} \quad , \qquad (5)$$

while the impulse response $h(t)$ may be expressed as a weighted sum of exponentials,

$$h(t) = \begin{cases} \sum_{k=1}^{N} a_k e^{s_k t} , & t \geq 0 \\ 0 , & t < 0 \end{cases} \qquad (6)$$

where the a_k and s_k are to be chosen to minimize the index I_2. The time domain problem, then, is to approximate the delayed sin x/x function which exists for all time by a weighted sum of exponentials which is nonzero only in positive time. The error committed in not approximating the tail of the $\sin (t - t_o)/(t - t_o)$ function in negative time is, of course, included in the total error, and in most cases constitutes the major contribution to I_2.

McDonough has recently published a review of the problem of least-squares approximation in positive time together with new results and a bibliography.(1) A compact statement of this problem is obtained by differentiating I_2 with respect to the a_k and s_k and setting each expression equal to zero.

$$\sum_{j=1}^{N} \frac{-1}{(s_j + s_k)} a_j = F(-s_k) \qquad (\partial I_2/\partial a_k = 0) \qquad (7)$$

$$k = 1,\ldots,N$$

$$\sum_{j=1}^{N} \frac{-1}{(s_j + s_k)^2} a_j = \frac{d}{ds}\Big[F(s)\Big]_{s = -s_k} \qquad (\partial I_2/\partial s_k = 0) \qquad (8)$$

The 2N equations (7) and (8) taken together are known as the equations of Aigrain and Williams(2) and must be simultaneously satisfied to minimize I_2. Both sets are nonlinear in the s_k, and have not been found to be a fruitful starting point in an investigation of optimum pole positions.

A better approach seems to be the use of a finite number of realizable rational functions $\Phi_j(s)$ selected from an orthonormal set in $L^2(-\infty,\infty)$; i.e.,

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_j(j\omega)\, \Phi_k(j\omega)\, d\omega \;=\; \delta_{jk} \quad . \tag{9}$$

The corresponding time functions $\phi_j(t)$, the inverse Fourier transforms of the functions $\Phi_j(s)$, also form an orthonormal set in $L^2(0,\infty)$. The time functions can be made complete in $L^2(0,\infty)$ by proper choice of the remaining pole positions, (3) but since only a finite number of terms can be used completeness need not be considered. We may approximate the ideal lowpass in either the time domain or frequency domain as a weighted sum of orthonormal functions;

$$f(t) \cong \sum_{k=1}^{N} b_k \phi_k(t) \overset{\Delta}{=} h(t) \quad , \tag{10}$$

or

$$F(j\omega) \cong \sum_{k=1}^{N} b_k \Phi_k(j\omega) \overset{\Delta}{=} H(j\omega) \quad . \tag{11}$$

The completeness of the time functions $\phi_j(t)$ would only guarantee that the integral squared error involved in the approximation (10) in positive time could be made arbitrarily small.

It can easily be shown that the weights which minimize the indices of performance I_1 and I_2 are identical and are given by(4)

$$b_k = \int_0^{\infty} f(t)\,\phi_k(t)\, dt \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\omega)\,\Phi_k^*(j\omega)\, d\omega \quad . \tag{12}$$

Using the relation (12) for the weights the indices may be calculated as

$$I_1 \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(j\omega)|^2\, d\omega - \sum_{k=1}^{N} |b_k|^2 \tag{13}$$

and

$$I_2 = \int_{-\infty}^{\infty} |f(t)|^2 \, dt - \sum_{k=1}^{N} |b_k|^2 \quad . \tag{14}$$

Because of the nature of the ideal lowpass both the weights and the index are most easily calculated by frequency domain integration. For the ideal lowpass equations (12) and (14) may be rewritten

$$b_k = \frac{1}{2\pi} \int_{-1}^{1} \Phi_k^*(j\omega) \, e^{-j\omega t_o} \, d\omega = \frac{1}{\pi} \mathrm{Re}\left\{ \int_{0}^{1} \Phi_k(j\omega) \, e^{j\omega t_o} \, d\omega \right\} \tag{15}$$

$$I_1 = \frac{1}{\pi} - \sum_{k=1}^{N} |b_k|^2 . \tag{16}$$

Happily, the required set of realizable rational functions does exist and was originally described by Kautz.(5) These functions, as modified by Young,(6) have the following description:

$$\Phi_{2k-1}(s) = \frac{\sqrt{-2p_k} \quad s}{s^2 - p_k s + q_k} \prod_{j=1}^{k-1} \frac{s^2 + p_j s + q_j}{s^2 - p_j s + q_j}$$

$$\Phi_{2k}(s) = \frac{\sqrt{-2p_k \, q_k}}{s^2 - p_k s + q_k} \prod_{j=1}^{k-1} \frac{s^2 + p_j s + q_j}{s^2 - p_j s + q_j}$$

$$k = 1,2,\ldots, \begin{cases} \frac{N-1}{2} & , \ N \ \text{odd} \\ N/2 & , \ N \ \text{even} \end{cases} \tag{17}$$

$$\Phi_N(s) = \sqrt{\frac{-2s_N}{s - s_N}} \prod_{j=1}^{\frac{N-1}{2}} \frac{s^2 + p_j s + q_j}{s^2 - p_j s + q_j} \quad , \ N \ \text{odd} \ ,$$

where

$$s^2 - p_k s + q_k \overset{\Delta}{=} (s - s_{2k-1})\,(s - s_{2k})$$

and either s_{2k-1}, s_{2k} are real, or $s_{2k} = s^{*}_{2k-1}$. The problem may now be stated in terms of a choice of the best pole positions of the $\Phi_j(s)$, since

$$I_1 = \frac{1}{\pi} - \sum_{j=1}^{N} b_j^2 = \frac{1}{\pi}\left[1 - \frac{1}{\pi}\sum_{j=1}^{N} \mathrm{Re}^2\left\{\int_0^1 \Phi_j(j\omega)\, e^{j\omega t_o}\, d\omega\right\}\right]. \tag{18}$$

Once these pole positions are found, the b_j are determined through use of equation (15) and finally $H(s)$ itself through equation (11).

III. NUMERICAL DETERMINATION OF THE OPTIMUM POLE POSITIONS

In equation (18) the index to be minimized has been expressed in terms of the pole positions of the approximating filter alone. There being no known analytical technique to solve this problem, the minimization was carried out with the aid of a digital computer. One great advantage of the orthonormal frequency representation immediately appears - the integration required to obtain the b_j extends over a finite interval whenever the function to be approximated is band-limited.

The principal difficulty involved in locating the position of the minimum of I_1 in the space spanned by the p_k and q_k is that I_1 is relatively insensitive to these parameters.(1) This fact precluded use of "direct search" techniques where action is taken only on the basis of information gained by evaluating I_1 at some set of points. Since the partial derivatives $\partial I_1/\partial p_k$ and $\partial I_1/\partial q_k$ can be evaluated without much difficulty, a "steepest descents" procedure was attempted, in which a new guess is formed from the old by adding to p_k and q_k the quantities

$$- \Delta \frac{\partial I_1/\partial p_k}{\gamma} \quad \text{and} \quad - \Delta \frac{\partial I_1/\partial q_k}{\gamma}$$

respectively, where

$$\gamma \overset{\Delta}{=} \left(\sum_k \left[\partial I_1/\partial p_k^2 + \partial I_1/\partial q_k^2 \right] \right)^{\frac{1}{2}} \quad , \qquad (19)$$

and Δ is a predetermined step size which is reduced as I_1 approaches its minimum. While this method worked well to bring an initial guess of the p_k and q_k to the general vicinity of the minimum, its convergence was so slow as to make the method useless for obtaining final results. Another attempt made use of McDonough's suggestion to minimize γ rather than I_1 itself, the advantage being that it is far more sensitive to the parameters and in fact is precisely zero at the stationary point.(1) Its principal disadvantage is the presence of saddle points quite close to the minimum desired. The steepest descent technique applied to γ required the evaluation of the second partial derivatives of I_1, a far more complicated chore than that of the first derivatives. With this task accomplished, it was found that the new method converged no more quickly than the old. The next approach was to apply the so-called Newton's method, where the approximate distances δ_{p_k} and δ_{q_k} from a given point P to the stationary point are found by solving the matrix equation

$$L \; d = - \Delta \quad , \qquad (20)$$

in which L is the N x N matrix of second derivatives having components

$$\left\{\partial^2 I_1/\partial p_i\ \partial p_j,\ \partial^2 I_1/\partial p_i\ \partial q_j\ \text{ and }\ \partial^2 I_1/\partial q_i\ \partial q_j\right\} ,$$

$$d \triangleq \operatorname{col}\left[\delta p_1 \ldots \delta p_k \ldots \delta q_1 \ldots \delta q_k \ldots\right] , \tag{21}$$

and

$$\Delta \triangleq \operatorname{col}\left[\partial I_1/\partial p_1 \ldots \partial I_1/\partial p_k \ldots \partial I_1/\partial q_1 \ldots \partial I_1/\partial q_k \ldots\right] , \tag{22}$$

where all partial derivatives are evaluated at P.(7) The distances are exact if the I_1 surface is a paraboloid, and the method will converge if repeated provided the matrix of second derivatives remains positive definite. In the case at hand, the δp_k and δq_k were used only to establish a direction. A search in a single variable (the proper distance) was then carried out to determine (roughly) how far one should go in this direction before the line of steepest descent is at right angles to the present direction. The method used for this search is due to Kiefer and is reported in Spang.(7) An illustration of the scheme is shown in Figure 2. The overall search was repeated until Y fell below a preset cutoff. The method turned out to converge quickly if the initial guess was good enough (five or six matrix inversions brought Y below 10^{-7}). Divergence set in immediately otherwise. The final procedure was as follows:

(a) For a new N, choose as initial pole positions the poles of the corresponding Butterworth filter and a t_o of about four times that of the Butterworth. Perform a steepest descents search on I_1 until Δ becomes less than about .001.

(b) Using the results of (a), locate the exact stationary point by Newton's method.

(c) Reduce t_o and repeat (b) using the previous results as starting values for the new minimization. There is a limit as to the amount t_o can change between minimizations before divergence sets in.

(d) After the best pole positions for different t_o have been determined, repeat (c) including t_o as a parameter, using as initial guesses those parameters p_k, q_k and t_o obtained in (c) for every case in which the quantity $\partial I_1/\partial t_o$ was near zero and $\partial^2 I_1/\partial t_o^2$ positive. This yields several filters for each N which possess an I_1 which is minimum with respect to the parameter t_o.

Since the function $F(j\omega)$ could be evaluated for any desired value of ω, the required integrations were performed using Gauss Quadrature. It was soon discovered that one of the poles of the least-squares filter is very close to the $j\omega$ axis near $\omega = 1$, causing a large peak in the magnitude function of any $\Phi_j(s)$ containing it. To combat the serious integration error which would otherwise arise, this pole was assigned the largest value of k, so that at most 3 $\Phi_j(s)$ contained it. In addition, the integration was performed in two parts;

$$\int_0^{1 + 2\mathrm{Re}\left\{s_k\right\}} d\omega \quad \text{and} \quad \int_{1 + 2\mathrm{Re}\ s_k}^{1} d\omega.$$

The entire peak is contained in the second of these intervals.

Once the optimum pole positions have been found, calculation of any other desired quantity is straightforward. H(s) itself may be calculated as a quotient of polynomials by collecting the sum expressed in equation (11) over a common denominator. The time response may then be found by performing a partial fraction expansion to obtain the a_j in equation (6).(8)

IV. RESULTS

The pole positions of the optimum least-squares filters obtained in step (d) of Section III are listed in Table 1 for $N = 4,6,8$, and 10. Not enough points were used in the quadrature integration formula to provide better than 3 digit accuracy. The corresponding approximation error, expressed as $100 \pi I_1$ %, is also given. The remarkable fact about the pole positions of these filters is that for $N > 4$ the filters possess multiple poles. Just how the multiplicity changes for larger N or for odd N has not yet been determined, nor is a proof of this phenomenon available. The fact remains true whether t_o is prespecified or is a parameter in the minimization. The loci of the poles of the eighth order filter with delay t_o a parameter are shown in Figure 3.

The frequency response of one of the sixth order filters is shown in Figure 4. Since it has two less zeros than poles, it attenuates eventually at a rate of 12 db./octave. Note that both the magnitude and phase are being approximated by this method. In no case so far has the peak in the last $\Phi_j(j\omega)$ carried over to $H(j\omega)$, due to a nearby zero in H(s).

Another interesting phenomenon turned up when the time response of the filters found in III (d) were evaluated. In all cases where t_o turned out such that $f(0) = \sin t_o/t_o$ was positive, $H(s)$ has two less zeros than poles and therefore $h(0) = 0$. Where $f(0)$ was negative, $H(s)$ has only one less zero than the number of poles, and $h(0) = 2 \sin t_o/t_o$. The error at $t = 0$ is thus in all cases tested equal to $|\sin t_o/t_o|$. No explanation of this behavior can be offered at present. The time response of the sixth order filter of Figure 4 is shown in Figure 5.

To illustrate how little the index depends on the pole positions, the parameters of the sixth order filter with Butterworth poles but zeros chosen according to the least-squares criterion are presented in Table 1 and the frequency response given in Figure 6. The error is 1.07% higher than the corresponding error of the optimum least-squares filter.

V. CONCLUSIONS

The methods by which the parameters of a new class of filter may be obtained has been described and sample results given. Although the technique is essentially in its final form, the data given must be regarded as preliminary. It is hoped that eventually a catalog of these filters may be published involving more precision in the pole position calculation. A number of extensions will be made as soon as possible. The first will be to obtain the approximation to any band-limited signal, followed by a modification to handle time-limited signals. The last extension for analytically described signals will be to functions of unlimited extent in time and frequency. A more difficult problem is the approximation of signals in sampled-data form, but the same general techniques should apply.

The question might well arise as to the value of carrying out the research reported in this paper, or any of the extensions mentioned above. Since almost any choice of poles give nearly as good results, why bother with a time-consuming minimization procedure? The answer we feel is twofold. In the first place, the problem is an esthetic one - it has existed for a long time with very little known about its solution. The second reason is that the problem seemed admirably suited to serve as a starting point for further theoretical and applied research in signal theory and signal processing techniques. We believe the startling results have already justified the time spent.

VI. REFERENCES

(1) R. N. McDonough, Representation and Analysis of Signals; Part XV, Matched Exponents for the Representation of Signals. Dept. of Electrical Engineering, The Johns Hopkins University, April 1963.

(2) P. R. Aigrain and E. M. Williams, "Synthesis of n-Reactance Networks for Desired Transient Response," J. Appl. Physics, Vol. 20, pp. 597-600 (June 1949).

(3) R. E. A. C. Paley and N. Wiener, "Fourier Transforms in the Complex Domain," Amer. Math. Soc. Coll. Pub., Vol. XIX, New York(1934) (Szasz's Theorem).

(4) R. Courant and D. Hilbert, Methods of Mathematical Physics. Vol. I, Interscience, New York, 1953, p. 52.

(5) W. H. Kautz, "Transient Synthesis in the Time Domain," IRE Trans. on Circuit Theory, Vol. CT-1 (September 1954), pp. 29-39.

(6) T. Y. Young, Representation and Analysis of Signals; Part X, Signal Theory and Electrocardiography. Dept. of Electrical Engineering, The Johns Hopkins University, May 1962.

(7) H. A. Spang, III, "A Review of Minimization Techniques for Nonlinear Functions," SIAM Review, Vol. 4 (October 1962), pp. 343-365.

(8) C. Pottle, "On the Partial Fraction Expansion of a Rational Function with Multiple Poles by Digital Computer," I.E.E.E. Trans. on Circuit Theory, Vol. CT-11 (March 1964), pp. 161-162.

TABLE 1

			Pole Locations			Zero Locations	
N	t_o	I min in %	Real Part	Imag. Part	Mult.	Real Part	Imag. Part
4	2.8923	5.60	- .4137 - .1087	.4228 .9664	1 1	- .1501	1.2258
4	5.4174	4.40	- .2970 - .1555	.3230 .8742	1 1	.5625 - .0927	- 1.6667
6	3.0078	5.18	- .6803 - .0485	.7042 .9958	2 1	- .4311 - .0618	1.8963 1.0667
6	5.8739	3.12	- .5449 - .0873	.5017 .9744	2 1	.5099 - .0661 - .2299	- 1.1514 3.0127
6	8.4022	2.96	- .4378 - .1206	.4097 .9216	2 1	.4128 - .0463	.3544 1.2948
8	6.1167	2.61	- .5883 - .1698 - .0149	.4517 .9292 .9996	2 1 1	.4971 - .0205 - .1780 - .8662	- 1.0175 1.2692 4.0797
8	9.0804	1.89	- .4822 - .2050 - .0307	.3543 .8246 .9950	2 1 1	.3623 - .1861 - .0327	.3532 1.6311 1.0468
8	11.868	1.69	- .3883 - .1871 - .0534	.3011 .7385 .9783	2 1 1	.3173 .2912 - .1166 - .0396	- .5219 2.5509 1.1053
10	12.138	1.45	- .5557 - .1810 - .0246	.4061 .8543 .9964	3 1 1	.3042 .2813 - .1857 - .0997	- .5210 3.9810 1.3926
10	14.900	1.36	- .4723 - .1758 - .0425	.3545 .7855 .9859	3 1 1	.2696 .2384 - .0666 - .0272	.2047 .6197 1.7087 1.0728
6 Butterworth Poles	3.0078	6.25	- .9659 - .7071 - .2589	.2589 .7071 .9659	1 1 1	- 1 .3756 - 5 .5308 - .1794	 1.5561 1.3957

$$F(j\omega) = \begin{cases} e^{j\omega t_0} & , |\omega|<1 \\ 0 & , |\omega|>1 \end{cases}$$

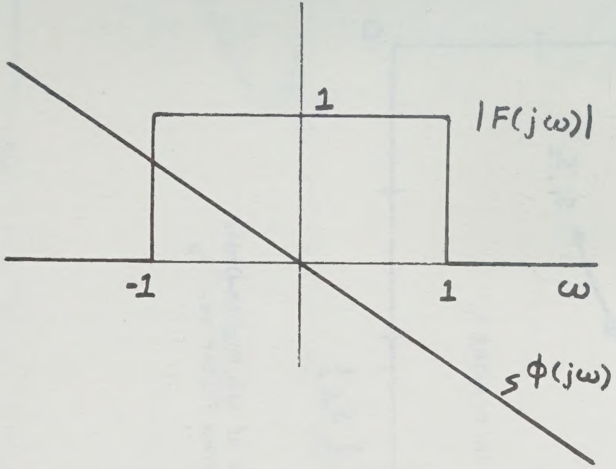


Fig. 1 - The Ideal Lowpass Filter

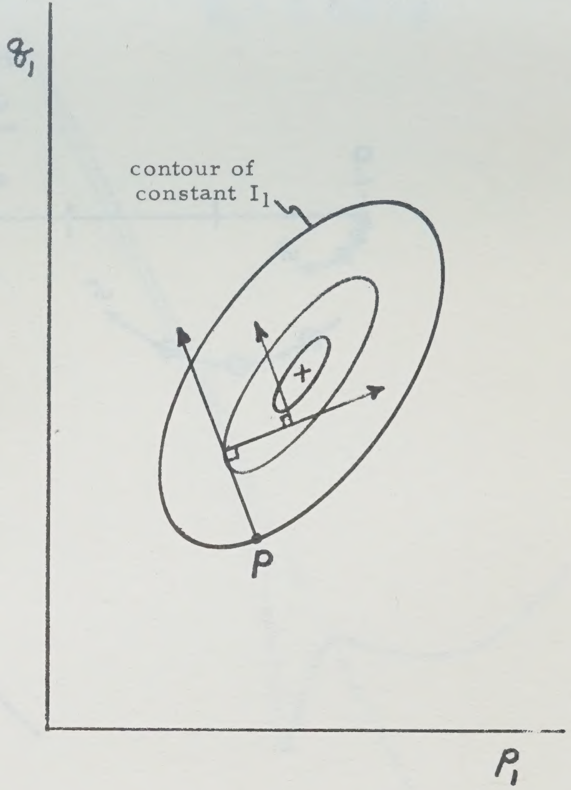

q₁
contour of
constant I₁
P
p₁

Fig. 2 - An Illustration of Newton's Method

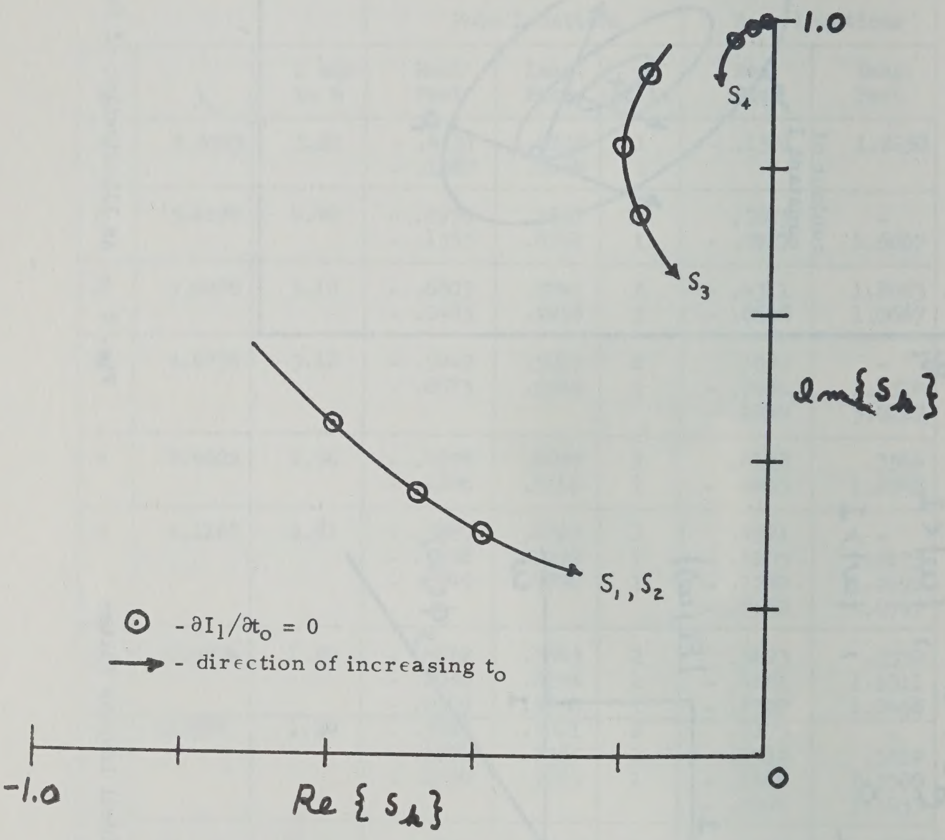


Fig. 3 - Pole Locus of the Eighth-Order Least-Squares Filter vs. t_o

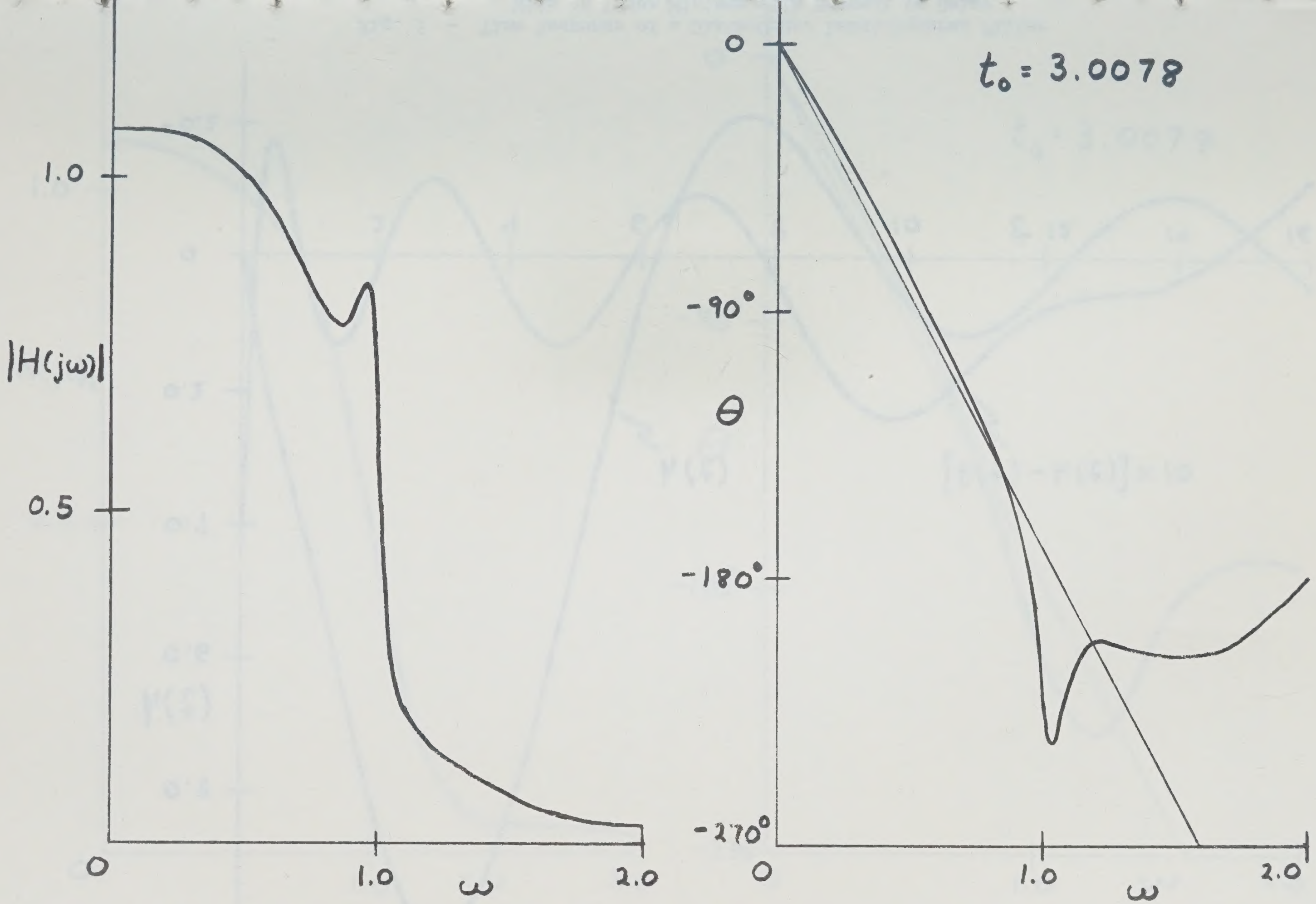
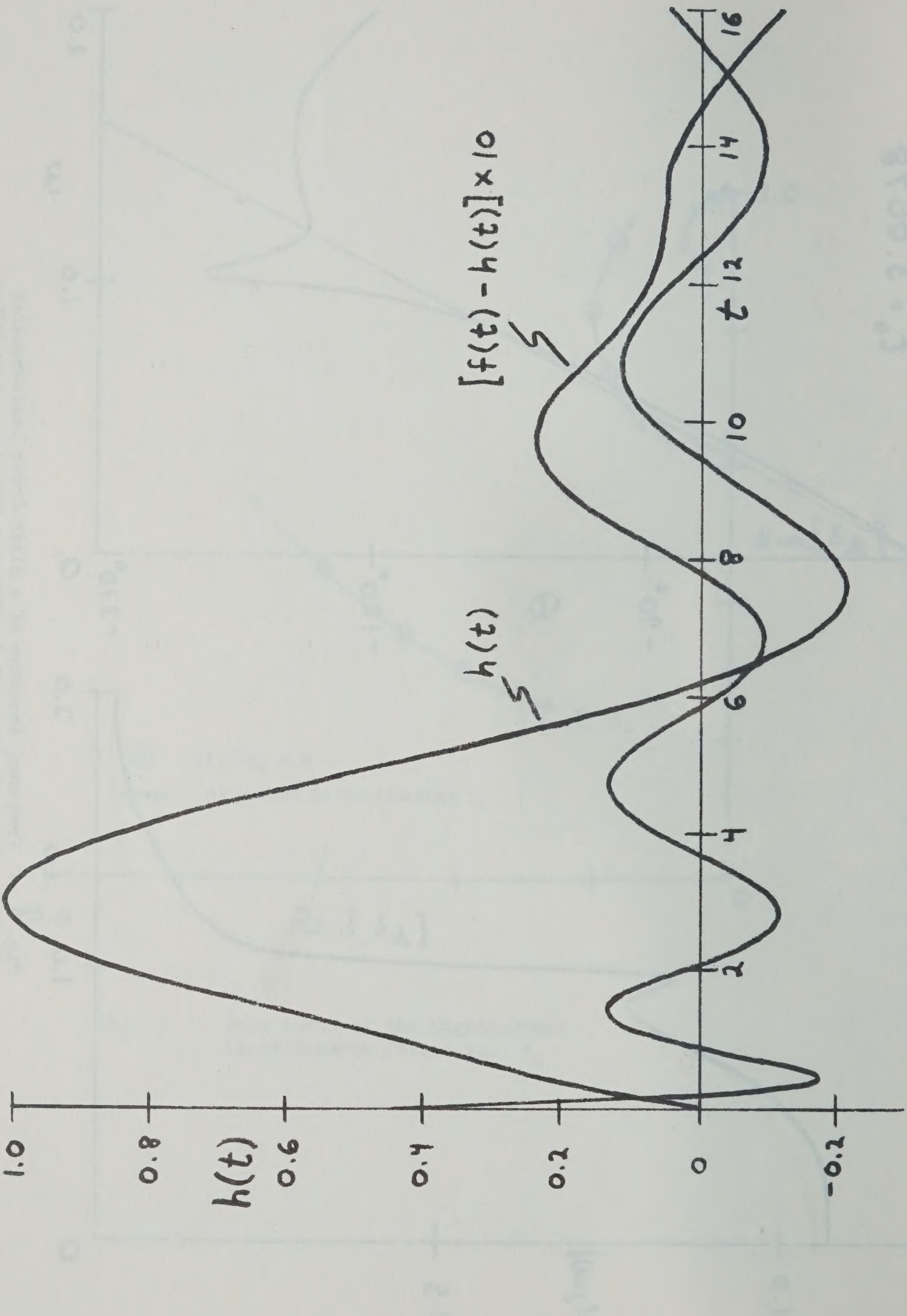


Fig. 4 - Frequency Response of a Sixth-Order Least-Squares Filter with an Index Minimum with Respect to Delay


h(t)
1.0
0.8
0.6
0.4
0.2
0
-0.2
t
2
4
6
8
10
12
14
16
h(t)
[f(t) - h(t)] × 10

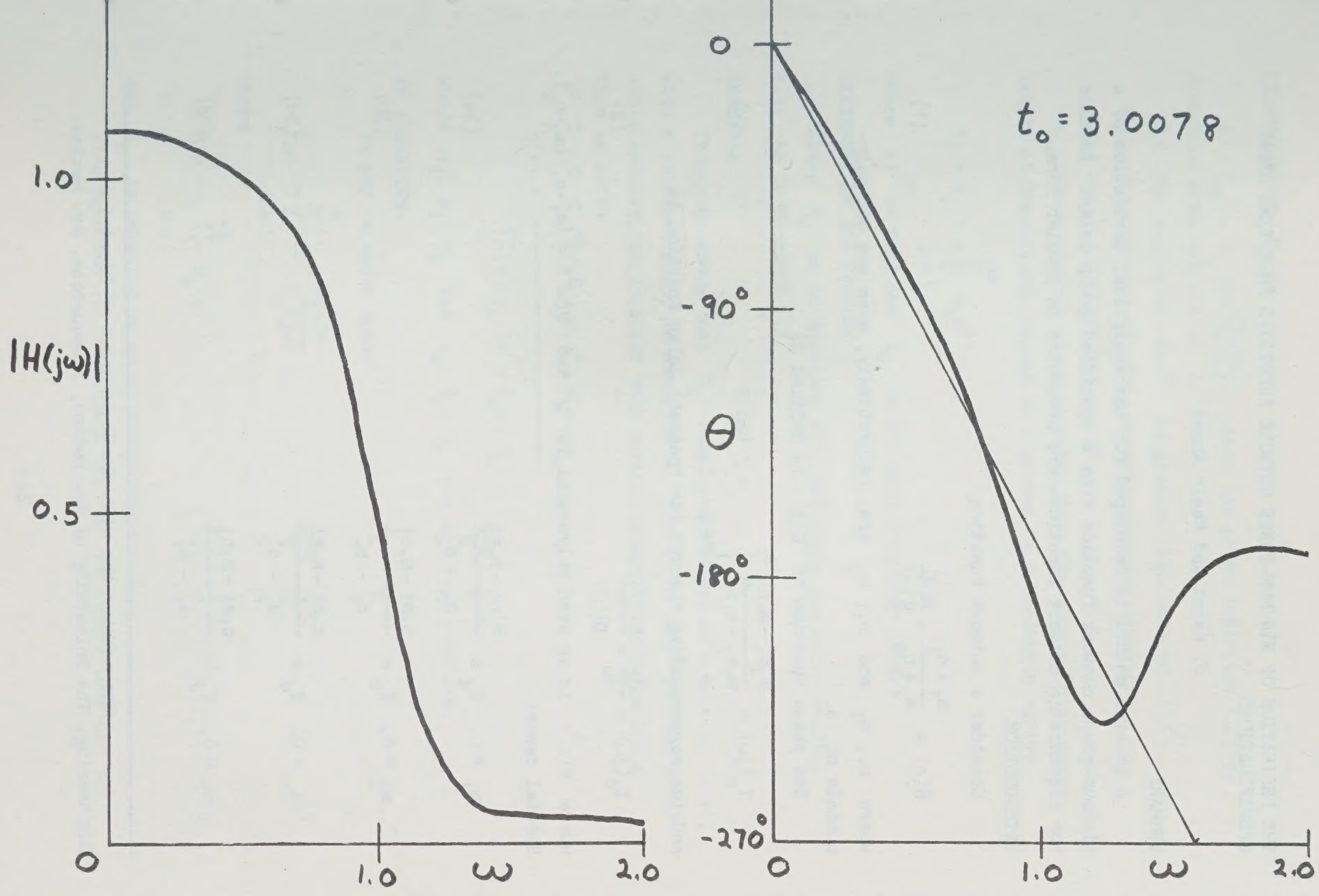


Fig. 6 - Frequency Response of a Sixth-Order Filter with Butterworth Poles and Least-Square Zeros with the Same Delay as in the Previous Cases

THE DERIVATION OF MINIMUM-PHASE NETWORK FUNCTIONS FROM TIME DELAY SPECIFICATIONS

S. Karni and Nasir Ahmed

SUMMARY

A general method is developed for the analytical derivation of a minimum-phase network function from a specified group delay. Results for biquadratic network functions are presented in tabular form.

INTRODUCTION

Consider a network function

$$H(s) = \frac{m_1 + n_1}{m_2 + n_2} = \frac{p(s)}{q(s)} \tag{1}$$

where m_1, m_2 and n_1, n_2 are, respectively, even and odd polynomials of s.

The phase function of $H(s)$ is defined as

$$T_p(j\omega) = \left. \frac{m_2 n_1 - m_1 n_2}{m_1 m_2 - n_1 n_2} \right]_{s = j\omega} = j \tan \beta \tag{2}$$

and the corresponding "group" (or "phase") delay function is

$$T_d(j\omega) = \frac{-d\beta}{d\omega} = \frac{N(\omega^2)}{D(\omega^2)} \tag{3}$$

where $N(\omega^2)$ is an even polynomial in ω and $D(\omega^2) = (m_1^2 - n_1^2)(m_2^2 - n_2^2)$. Special cases:

$$m_1 \equiv 0, \quad T_d = \frac{m_2 n_2' - n_2 m_2'}{m_2^2 - n_2^2} \tag{3a}$$

$$m_2 \equiv 0, \quad T_d = \frac{n_1 m_1' - m_1 n_1'}{m_1^2 - n_1^2} \tag{3b}$$

$$n_1 \equiv 0, \quad T_d = \frac{m_2 n_2' - n_2 m_2'}{m_2^2 - n_2^2} \tag{3c}$$

$$n_2 \equiv 0, \quad T_d = \frac{n_1 m_1' - m_1 n_1'}{m_1^2 - n_1^2} . \tag{3d}$$

The authors are with the Department of Electrical Engineering, The University of New Mexico, Albuquerque, New Mexico.

PROPERTIES OF $T_d(j\omega)$

i) $T_d(j\omega) = -\text{Re}[\frac{H'(j\omega)}{H(j\omega)}]$ where the prime indicates differentiation with respect to s.

ii) The degree of $D(\omega^2)$ is greater than that of $N(\omega^2)$.

iii) The poles of $T_d(j\omega)$ possess quadrant symmetry.

iv) $D(\omega^2) > 0$ for $0 < \omega < \infty$.

v) Since $T_d(j\omega)$ is a rational even function with no singularity at the origin, it may be expanded in a Maclaurin series:

$$T_d = t_o(1 + \sum_{i=1}^{\infty} a_i\omega^{2i})$$

where a_i is real and t_o is the zero-frequency delay.

STATEMENT OF THE PROBLEM

Given T_d as in Equation (3), determine the corresponding $H(s)$ as in Equation (1).

ANALYSIS

It may be shown that T_d remains unchanged if a zero of $q(s)$, with a reversed sign, is assigned to $p(s)$. In order to remove this arbitrariness, we consider only minimum-phase network functions. Thus we write

$$H(s) = \frac{\prod_i (s+\alpha_i)(s^2+\beta_i s+\gamma_i)}{\prod_k (s+\rho_k)(s^2+\xi_k s+\eta_k)} \tag{4}$$

where α_i, β_i, γ_i and ρ_k, ξ_k, η_k are all real positive.

AN ALGORITHM

It may be shown that:(1)

$$T_d = -\sum_{i=1}^{N} \text{Re}[\frac{H_i'(j\omega)}{H_i(j\omega)}] \tag{5}$$

where

$$H(s) = \prod_{k=1}^{N} H_k(s) \tag{6}$$

and H_k is referred to as an auxiliary function.

From Equation (4) it follows that we can always consider the auxiliary functions to be first order or second order polynomials or their reciprocals, with real positive coefficients. We list these with their corresponding auxiliary time delay functions T_{dk} in Table 1.

Table 1

H_k	$T_{dk}(\alpha > 0,\ \beta > 0)$
$s+\alpha$	$\frac{-\alpha}{\omega^2+\alpha^2}$
$\frac{1}{s+\alpha}$	$\frac{+\alpha}{\omega^2+\alpha^2}$
$s^2+\alpha s+\beta$	$\frac{-\alpha(\beta+\omega^2)}{\omega^4+(\alpha^2-2\beta)\omega^2+\beta^2}$
$\frac{1}{s^2+\alpha s+\beta}$	$\frac{\alpha(\beta+\omega^2)}{\omega^4+(\alpha^2-2\beta)\omega^2+\beta^2}$

Thus, a given T_d may always be expressed as:

$$T_d = \frac{N(\omega^2)}{D(\omega^2)} = \sum_{k=1}^{N} T_{dk} = \sum_{k=1}^{N} \delta_k \left[\frac{n_k(\omega^2)}{d_k(\omega^2)} \right] \tag{7}$$

where $\delta_k = \pm 1$, and $D(\omega^2) = \prod_{k=1}^{N} d_k(\omega^2)$. (7a)

OUTLINE OF THE METHOD

i) Factorize $D(\omega^2)$. Here the method of iterative factorization$^{(2)}$ may be used to good advantage.

ii) From Table 1 it follows that each $d_k(\omega^2)$ yields the corresponding T_{dk} within the arbitrariness of a positive or negative sign. As a result, Equation (7) can be written.

iii) Obtain a system of linear simultaneous equations for δ_k by matching the coefficients of $N(\omega^2)$ with the numerator obtained from the expansion $\sum_{k=1}^{N} \delta_k \left[\frac{n_k(\omega^2)}{d_k(\omega^2)} \right]$. Solve for δ_k.

iv) Having the sign of each T_{dk}, find the corresponding H_k, as shown in Table 1.

v) Obtain the desired network function $H(s) = \prod_{k=1}^{N} H_k(s)$.

THE ITERATIVE METHOD OF FACTORIZATION OF POLYNOMIALS

Here we merely list the steps necessary for the purposes of computation. The reader may refer to Carnahan et al. for a detailed discussion.

It is desired to factorize

$$f(x) = \sum_{i=0}^{n} a_i x^{n-1}, \; a_o = 1. \tag{8}$$

We consider $g(x) = \sum_{i=0}^{m} p_i x^{m-1}$, $p_o = 1$, $m < n$, to be the initial approximation. Therefore

$$f(x) = g(x)h(x) + \sum_{i=0}^{m-1} r_i x^i \tag{9}$$

where $\sum_{i=0}^{m-1} r_i x^i$ is the remainder term. Convergence is achieved as the remainder term diminishes.

The following set of equations is used for the iteration:

$$a_k = \sum_{i=0}^{m} p_i^{(\ell)} b_{k-i}^{(\ell)}, \; b_s^{(\ell)} = 0, \; s < 0, \; 0 \leq k \leq m. \tag{10}$$

Equation (10) yields the parameters b_k, which are substituted in Equation (11) to obtain the parameters c_k. Here ℓ represents the ℓ-th iteration.

$$-b_k^{(\ell)} = \sum_{i=0}^{m} p_i^{(\ell)} \; c_{k-i}^{(\ell)}, \; c_s^{(\ell)} = 0, \; s < 0, \; 0 \leq k \leq m. \tag{11}$$

Once b_k and c_k are known, Equation (12) is used to solve for the p_i's required for the $(\ell + 1)$-st iteration.

$$\sum_{i=1}^{m} p_i^{(\ell+1)} \; c_{k-i}^{(\ell)} = -2b_k^{(\ell)} - c_k^{(\ell)}, \; n-m+1 \leq k \leq n. \tag{12}$$

Convergence is not guaranteed, but may easily be checked for by using the condition

$$\sum_{i=n-m+1}^{n} |b_i| < \varepsilon, \tag{13}$$

where ε is an arbitrarily small positive number.

EXAMPLE I

$$T_d = \frac{-2\omega^2 + 69\omega^2 + 180}{\omega^6 + 28\omega^4 + 84\omega^2 + 225} . \qquad (14)$$

It is required to determine the corresponding H(s).

Solution:

$$D(\omega^2)\Big]_{\omega^2 = x} = f(x) = x^3 + 28x^2 + 84x + 225.$$

We seek $f(x) \approx (x^2 + p_1 x + p_2)(x + b_1)$, and from Equations (10), (11) and (12) obtain:

$$b_1 = a_1 - p_1 \qquad c_1 = -b_1 + p_1$$

$$b_2 = a_2 - b_1 p_1 - p_2 \qquad c_2 = -b_2 - c_1 p_1 + p_2$$

$$b_3 = a_3 - b_1 p_2 - b_2 p_1 \qquad c_3 = -b_3 - c_1 p_2 - c_2 p_1$$

$$p_1' c_1 - p_2' = -2b_2 - c_2$$

$$p_1' c_2 + p_2' c_1 = -2b_3 - c_3. \qquad (15)$$

The criterion for convergence is $|b_2| + |b_3| < \varepsilon$. Primed letters are used for the $(\ell + 1)$-st step and unprimed letters for the ℓ-th step. In this example, $a_1 = 28$, $a_2 = 84$, $a_3 = 225$.

With $g(x) = x^2 + 0x + 0$ we develop Table 2 in succession.

Table 2

	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$
p_1	0	3.309	3.102	3.000	3
p_2	0	-0.363	9.554	8.981	9
b_1	28	24.061	24.898	25	25
b_2	84	11.242	-2.788	0.019	0
b_3	225	199.57	-4.227	0.418	0
c_1	-28	-21.022	-21.796	-22	
c_2	-84	52.281	79.953	74.962	
c_3	-225	-350.631	-35.639	-27.724	
p_1'	3.309	3.102	3	3	
p_2'	-0.363	9.554	8.981	9	

Since $|b_2| + |b_3| = 0$ ($\ell = 5$), the iteration has converged to the exact solution. Thus

$$D(\omega^2)\Big]_{\omega^2 = x} = (x^2 + 3x + 9)(x + 25)$$

or

$$D(\omega^2) = (\omega^4 + 3\omega^2 + 9)(\omega^2 + 25);$$

i) $\omega^4 + 3\omega^2 + 9$ yields $\beta = 3$, and with $\alpha^2 - 2\beta = 3$ we find $\alpha = 3$.

ii) $\omega^2 + 25$ yields $\alpha = 5$.

Using the algorithm, Equation (7), we have

$$T_d = \frac{-2\omega^4 + 69\omega^2 + 180}{\omega^6 + 28\omega^4 + 84\omega^2 + 225} = \delta_1\{\frac{5}{\omega^2 + 25}\} + \delta_2\{\frac{3(\omega^2 + 3)}{\omega^4 + 3\omega^2 + 9}\}. \tag{16}$$

Matching coefficients of the numerators of Equation (16), we obtain

$$\begin{aligned} 5\delta_1 + 3\delta_2 &= -2 \\ 45\delta_1 + 225\delta_2 &= 180, \end{aligned} \tag{17}$$

therefore

$$\delta_1 = -1, \ \delta_2 = 1.$$

Thus $H_1 = \frac{(s+5)}{1}$; $H_2 = \frac{1}{s^2 + 3s + 3}$ and the desired network function is

$$H(s) = H_1 H_2 = \frac{s+5}{s^2 + 3s + 3} . \tag{18}$$

Similarly, if $T_d(j\omega) = \dfrac{8\omega^4 + 99\omega^2 + 270}{\omega^6 + 28\omega^4 + 84\omega^2 + 225}$ we would obtain $\delta_1 = \delta_2 = +1$. Then $H(s) = \dfrac{1}{(s+5)(s^2+3s+3)}$.

The network function derived by the above method is arbitrary within: (i) a constant multiplier and (ii) a lossless network, as these two do not contribute to the time delay function.

SIMPLIFICATION OF THE METHOD FOR BIQUADRATIC DRIVING-POINT FUNCTIONS

No factorization of $D(\omega^2)$ is required if the given time delay function corresponds to a biquadratic driving-point function. The results obtained are tabulated in Table A.

Consider

$$H(s) = \frac{s^2 + a_1 s + a_o}{s^2 + b_1 s + b_o} \tag{19}$$

for which

$$T_p(s) = \frac{\lambda_3 s^3 + \lambda_1 s}{s^4 + \lambda_2 s^2 + \lambda_o} = \frac{N(s)}{s^4 + \lambda_2 s^2 + \lambda_o} \tag{20}$$

where $\lambda_3 = a_1 - b_1$, $\lambda_2 = a_o + b_o - a_1 b_1$, $\lambda_1 = a_1 b_o - a_o b_1$, and $\lambda_o = a_o b_o$. All the a's and b's are real positive.

It can be shown that if

$$T_d = \frac{x_6 s^6 + x_4 s^4 + x_2 s^2 + x_o}{s^8 + \hat{x}_6 s^6 + \hat{x}_4 s^4 + \hat{x}_2 s^2 + \hat{x}_o} \tag{21}$$

then

$$\lambda_o = \sqrt{\hat{x}_o}\ , \ \lambda_3 = x_6,\ \lambda_1 = \frac{-x_o}{\sqrt{\hat{x}_o}},\ \lambda_2 = \frac{x_6^2 + \hat{x}_6}{2}. \tag{22}$$

Now, the necessary and sufficient condition for $H(s)$ to be positive-real is $\lambda_2 \leq 2\sqrt{\lambda_o}$. Thus, from a given T_d as in Equation (21) the corresponding $T(s)$ may be constructed using Equation (22); if $\lambda_2 \leq 2\sqrt{\lambda_o}$, the following method may be used:

Assume the product $m_1 m_2$ to be of the form $(s^2+\theta)(s^2+\hat{\theta})$ where θ and $\hat{\theta}$ are real positive and $\theta\hat{\theta} = \lambda_o$.

Consider the following partial fraction expansion:

$$\frac{\lambda_3 s^3 + \lambda_1 s}{(s^2+\theta)(s^2+\hat{\theta})} = \frac{2k_1 s}{(s^2+\theta)} + \frac{2k_2 s}{(s^2+\hat{\theta})} = \frac{n_1}{m_1} - \frac{n_2}{m_2}. \tag{23}$$

From Equation (23) it follows that the signs of $2k_1$ and $2k_2$ will enable us to separate $\frac{n_1}{m_1}$ in this partial fraction expansion. It is found that

$$2k_1 = \frac{\lambda_1 - \lambda_3\theta}{(\hat{\theta} - \theta)}, \quad 2k_2 = \frac{\lambda_1 - \lambda_3\hat{\theta}}{(\theta - \hat{\theta})}, \quad \theta \neq \hat{\theta}. \tag{24}$$

The case $\theta = \hat{\theta}$ is treated separately and the results included in Table A.

From Equation (23) we get

$$n_1 n_2 = \frac{(\lambda_1 - \lambda_3\theta)(\lambda_1 - \lambda_3\hat{\theta})}{(\theta - \hat{\theta})^2} s^2. \tag{25}$$

Therefore the denominator of Equation (20) may be expressed as

$$s^4 + \lambda_2 s^2 + \lambda_o = m_1 m_2 - n_1 n_2 = (s^2 + \theta)(s^2 + \hat{\theta}) - \frac{(\lambda_1 - \lambda_3\theta)(\lambda_1 - \lambda_3\hat{\theta})}{(\theta - \hat{\theta})^2} s^2. \tag{26}$$

Setting $\psi = \theta + \hat{\theta}$ and $(\theta - \hat{\theta})^2 = \psi^2 - 4\theta\hat{\theta} = \psi^2 - 4\lambda_o$ in Equation (26) and rearranging terms we obtain the following equation, which will be called the ψ-Equation.

$$\psi^3 - \lambda_2\psi^2 + (\lambda_1\lambda_3 - 4\lambda_o)\psi + (4\lambda_o\lambda_2 - \lambda_3^2\lambda_o - \lambda_1^2) = 0. \tag{27}$$

By direct substitution it may be shown that $\psi = \theta + \hat{\theta} = a_o + b_o$ is a root of the ψ-Equation.

CLASSIFICATION OF THE ψ-EQUATION

a) $\lambda_2 \leq 0$.

By using Descartes' rule of signs we find that the upper bound of real positive roots of the ψ-Equation is one. But it has been shown that the ψ-Equation has at least one real positive root $\psi = a_o + b_o$. We conclude that the ψ-Equation has exactly one real positive root when $\lambda_2 \leq 0$, and this root is sufficient to determine m_1, m_2, n_1, n_2 and hence $H(s)$.

b) $\lambda_2 > 0$.

The application of Descartes' rule shows that the ψ-Equation has one or three real positive roots. Thus in this case it may be necessary to solve the cubic ψ-Equation. Each root will yield a set of m_1, m_2, n_1, and n_2. The set that gives the specified $T_p(s)$ must be chosen.

c) Miscellaneous:

i) $\lambda_2 = \lambda_3 \sqrt{\lambda_o}$, implies $\theta = \hat{\theta}$.

ii) $\lambda_2 = 2\sqrt{\lambda_o}$, implies $H(s)$ is a minimum dp function.

These cases are included in Table A.

d) $\lambda_2 > 2\sqrt{\lambda_o}$.

In this case $H(s)$ is a non positive-real biquadratic function, or does not belong to the set of biquadratic functions at all. Here, the general method using iterative factorization must be employed. A special case in this category is the all-pass function which is treated separately (see Table A).

AN ITERATIVE METHOD TO SOLVE FOR REAL ROOTS OF THE ψ-EQUATION

In the ψ-Equation we have the relations $(\theta - \hat{\theta})^2 = (a_o - b_o)^2 = \psi^2 - 4\lambda_o$, where $\psi = \theta + \hat{\theta} = a_o + b_o > 0$. Since θ, $\hat{\theta}$ (or a_o, b_o) are all real, we require $\psi^2 > 4\lambda_o$ or $\psi > 2\sqrt{\lambda_o}$. Thus $\psi = 2\sqrt{\lambda_o}$ may be used to good advantage as the first approximation in our iterative procedure. We use the iterative relation$^{(2)}$

$$\psi^{(k+1)} = \frac{\psi_1^{(1)} f(\psi_2^{(k)}) - \psi_2^{(1)} f(\psi_1^{(k)})}{f(\psi_2^{(k)}) - f(\psi_1^{(1)})} \tag{28}$$

where $\psi_1^{(k)}$ and $\psi_2^{(k)}$ are two numbers such that $f(\psi_1^{(k)})$ and $f(\psi_2^{(k)})$ are opposite in sign and k is the number of the iteration.

EXAMPLE II

$$T_d(s) = \frac{-2s^6 - 16s^4 - 38s^2 - 336}{s^8 - 30s^6 + 273s^4 - 820s^2 + 576}. \tag{29}$$

It is desired to determine $H(s)$. Here

$x_6 = -2,\ x_4 = -16,\ x_2 = -38,\ x_o = -336,$

$\hat{x}_6 = -30,\ \hat{x}_4 = 273,\ \hat{x}_2 = 820,\ \hat{x}_o = 576.$

From Equation (22) we obtain $\lambda_o = 24$, $\lambda_3 = -2$, $\lambda_1 = \lambda 4$, $\lambda_2 = -13$. Further, $\lambda_2 < 2\sqrt{\lambda_o}$ which means that the ψ-Equation may be used. The ψ-Equation is:

$$\psi^3 + 13\psi^2 - 124\psi - 1540 = 0. \tag{30}$$

The first approximation is $\psi^{(0)} = 2\sqrt{\lambda_o} \approx 9{\cdot}8$. Using Equation (28) we obtain Table 3.

Table 3

	k = 1	k = 2	k = 3	k = 4
$\psi_1^{(k)}$	9.8	10.865	10.988	11
$\psi_2^{(k)}$	12	12	12	12
$f(\psi_1^{(k)})$	-536.88	-70.044	-6.297	0
$f(\psi_2^{(k)})$	572	572	572	572

The fourth iteration has converged to the exact solution

$$\psi = \theta + \hat{\theta} = 11. \tag{28}$$

Thus $(\theta - \hat{\theta})^2 = \psi^2 - 4\lambda_o = 25$ and

$$\theta - \hat{\theta} = \pm\, 5. \tag{29}$$

1) Consider $\theta + \hat{\theta} = 11$ and $\theta - \hat{\theta} = 5$. This yields $\theta = 8$, $\hat{\theta} = 3$. From Equation A.3 (Table A) we have

$$\frac{n_1}{m_1} - \frac{n_2}{m_2} = \left[\frac{\lambda_1 - \lambda_3\theta}{\hat{\theta} - \theta} \right] \frac{s}{s^2 + \theta} + \left[\frac{\lambda_1 - \lambda_3\hat{\theta}}{\theta - \hat{\theta}} \right] \frac{s}{s^2 + \hat{\theta}} \, . \tag{30}$$

Substituting for θ, $\hat{\theta}$ and λ's in Equation (30) we get

$$\frac{n_1}{m_1} - \frac{n_2}{m_2} = - \frac{6s}{s^2 + 8} + \frac{4s}{s^2 + 3} \, . \tag{31}$$

Therefore $m_1 = s^2 + 3$, $m_2 = s^2 + 8$, $n_1 = 4s$, $n_2 = 6s$, and

$H(s) = \dfrac{s^2 + 4s + 3}{s^2 + 6s + 8}$, a dp function.

2) Consider $\theta + \hat{\theta} = 11$ and $\theta - \hat{\theta} = -5$. We get $\theta = 3$, $\hat{\theta} = 8$. However, H(s) remains unchanged. This is so because the roles of θ and $\hat{\theta}$ are merely interchanged in Equation (30). That is, the solution is unique within an arbitrary constant multiplier irrespective of the sign chosen in the relation $\theta - \hat{\theta} = \pm \sqrt{\psi^2 - 4\lambda_o}$.

Table A

I. $$T_d(s) = \frac{x_6 s^6 + x_4 s^4 + x_2 s^2 + x_o}{s^8 + \hat{x}_6 s^6 + \hat{x}_4 s^4 + \hat{x}_o}\,, \quad \hat{x}_o > 0$$

$$T_p(s) = \frac{\lambda_3 s^3 + \lambda_1 s}{s^4 + \lambda_2 s^2 + \lambda_o}\,, \ \lambda_o > 0, \text{ where } \ \lambda_o = \sqrt{\hat{x}_o}\,, \ \lambda_3 = x_6,$$

$$\lambda_1 = -\frac{x_o}{\sqrt{\hat{x}_o}}\,, \ \lambda_2 = \frac{x_6^2 + \hat{x}_6}{2}\,.$$ If $\lambda_2 \leq 2\sqrt{\lambda_o}$, the following methods are applicable:

(a) $\lambda_2 \leq 0$ and $\lambda_1 \neq \lambda_3 \sqrt{\lambda_o}$; $\psi = \theta + \hat{\theta}$, $\theta - \hat{\theta} = \sqrt{\psi^2 - 4\lambda_o}$.

$$\psi^3 - \lambda_2 \psi^2 + (\lambda_1 \lambda_3 - 4\lambda_o)\psi + (4\lambda_o \lambda_2 - \lambda_3^2 \lambda_o - \lambda_1^2) = 0. \qquad \text{A.1}$$

Equation (A.1) has exactly one positive real root, $\psi = \theta + \hat{\theta}$ where $\hat{\theta} = \lambda_o / \theta$.

(b) $0 < \lambda_2 \leq 2\sqrt{\lambda_o}$, and $\lambda_1 \neq \lambda_3 \sqrt{\lambda_o}$.

Equation (A.1) has one or three positive real roots.

$$\frac{n_1}{m_1} - \frac{n_2}{m_2} = \left[\frac{\lambda_1 - \lambda_3 \theta}{\hat{\theta} - \theta} \right] \frac{s}{s^2 + \theta} + \left[\frac{\lambda_1 - \lambda_3 \hat{\theta}}{\theta - \hat{\theta}} \right] \frac{s}{s^2 + \hat{\theta}} \qquad \text{A.3}$$

where $\hat{\theta} = \frac{\lambda_o}{\theta} \neq \theta$.

Equation (A.3) yields $H(s) = \frac{s^2 + a_1 s + a_o}{s^2 + b_1 s + b_o}$ which is p.r.

$H(s)$ is a minimum function if $\lambda_2 = 2\sqrt{\lambda_o}$.

II. If $\lambda_1 = \lambda_3 \sqrt{\lambda_o}$ and $\lambda_2 \leq 2\sqrt{\lambda_o}$, then

$$H(s) = \frac{s^2 + a_1 s + a_o}{s^2 + b_1 s + b_o}\,, \text{ where } \ a_o = b_o = \sqrt{\lambda_o}\,, \qquad \text{A.4}$$

$$a_1 b_1 = 2\sqrt{\lambda_o} - \lambda_2 \qquad \text{A.5}$$

$$a_1 - b_1 = \lambda_3. \qquad \text{A.6}$$

Solving Equations (A.5) and (A.6) one obtains

$$a_1 = \tfrac{1}{2}\left[\lambda_3 + \sqrt{\lambda_3^2 + 4(2\sqrt{\lambda_o} - \lambda_2)} \right]$$

$$b_1 = \tfrac{1}{2}\left[\sqrt{\lambda_3^2 + 4(2\sqrt{\lambda_o} - \lambda_2)} - \lambda_3 \right].$$

III. If $\lambda_2 = 2\sqrt{\lambda_o} + \frac{\lambda_3^2}{4}$, $\sqrt{\lambda_o} = \lambda_1/\lambda_3$, where $\lambda_1, \lambda_3 < 0$, then

$H(s) = \frac{s^2 - as + b}{s^2 + as + b}$ $(a > 0,\ b > 0)$ which is an all-pass function, where $a = \frac{-\lambda_3}{2}$, $b = \sqrt{\lambda_o}$.

IV. $T_d(s) = \frac{x_4 s^4 + x_2 s^2 + x_o}{\hat{x}_6 s^6 + \hat{x}_4 s^4 + \hat{x}_2 s^2 + \hat{x}_o}$, $\hat{x}_o > 0$

$T_p(s) = \frac{\lambda_3 s^3 + \lambda_1 s}{\lambda_2 s^2 + \lambda_o}$, $\lambda_o > 0$,

where $\lambda_o = \sqrt{\hat{x}_o}$, $\lambda_1 = -\frac{x_o}{\lambda_o}$, $\lambda_2 = \frac{\lambda_1^2 + \hat{x}_2}{2\lambda_o}$, $\lambda_3 = -\frac{x_4}{\lambda_2}$.

If $\lambda_2 \leq 0$, then $H(s)$ is p.r. and the following method is applicable:

$$\lambda_3^2\theta^3 - \lambda_1\lambda_3\theta^2 + \lambda_2\lambda_o\theta - \lambda_o^2 = 0. \qquad \text{A.7}$$

Equation (A.7) has exactly one positive real root.

$$\frac{n_1}{m_1} - \frac{n_2}{m_2} = \left[\ \frac{\lambda_1 - \lambda_3\theta}{\theta}\ \right] \frac{s}{s^2 + \theta} + \frac{\lambda_3}{\hat{\theta}}\, s, \quad \hat{\theta} = \frac{\lambda_o}{\theta}. \qquad \text{A.8}$$

Equation (A.8) yields $H(s) = \frac{a_1 s + a_o}{s^2 + b_1 s + b_o}$ or its reciprocal.

BIBLIOGRAPHY

(1) Gouriet, G. G., "Two Theorems Concerning Group Delay with Practical Application to Delay Correction", Proc. IEE (England) part C, Vol. 105, pp. 240-244, March 1958.

(2) Carnahan, B., Luther, H. A., Wilkes, J. O., Applied Numerical Methods (Vol. 1), John Wiley and Sons Inc., pp. 194-198, 227-228.

(3) Mitra, S. K., "On the Construction of a Positive Real Impedance from a Given Even Part", Proc. IEEE, Vol. 51, pp. 1267, Sept. 1963.

This material is from the book Network Theory: Analysis and Synthesis by S. Karni, to be published by Allyn and Bacon, Inc., Boston, Mass.

ON THE CONSTRUCTION OF PHASE-TRAJECTORIES

G. C. Reis*

1. General Procedure

The most general procedure presently available for the phase plane solution of nonlinear equations is that due to Pell,(1,2) with a recent extension by Reis.(3) The basic philosophy of the method is explained in figure 1. At the starting point P, the slope of the trajectory is determined as follows. The equation for the slope is

$$\frac{d\dot{x}}{dx} = \frac{\ddot{x}}{\dot{x}} = - \frac{\phi(x,\dot{x})}{\dot{x}} \qquad (1)$$

where ϕ is, in general, a nonlinear function and the dot convention has been used to indicate differentiation. In figure 1, the point M is directly below P. The distance MN is set equal to $\phi(x, \dot{x})$. Since PM equals $\dot{x}$, the slope of PN is $\dot{x}/ \phi(x, \dot{x})$. Hence the line ab, constructed perpendicular to PN, has the slope $-\phi(x, \dot{x})/\dot{x}$, and is, therefore, the slope of the trajectory, as given by equation (1). Thus the trajectory can be constructed, provided that a convenient method for finding the distance MN is available. The next section deals with this problem.

2. Useful Graphical Manipulations

The most commonly encountered problem is: given the distances A, B and the unit distance, find the distance equal to the product AB. Figure 2 indicates the solution to this problem. At the end of length A, the unit distance is constructed perpendicular to A. A line from the free end of A to the free end of the unit distance is drawn. Two sides of the triangle thus formed are extended as shown in figure 2. The length B is then positioned between these extended lines, as shown. The distance x is then equal to the product AB. The proof of this follows directly from the similarity of the two right triangles.

The extension of this construction to products of more than two terms is direct. One merely lays off the unit distance on B, and proceeds to find xC, where C is the third factor, as shown in figure 3.

---

* Drexel Institute of Technology, Philadelphia, Pennsylvania

Another useful construction is to find a distance equal to the reciprocal of a given distance. The details are shown in figure 4. Once again the unit distance is laid off normal to the distance A at one end of A. The free ends of the two distances are joined to form a right triangle. The unit distance is then laid off on A, and the line x is drawn as shown. Now x equals 1/A as can be verified by the similarity of the two right triangles.

The final construction is that of extracting the square root, as shown in figure 5. In this figure the line segment OF is the length A. A line FG is constructed normal to OF. The unit length is constructed normal to OF at the point P. The line from O through the free end of the unit distance intersects FG at Q. The point P is adjusted until OP = FQ. Then $OP = QF = A^{\frac{1}{2}}$ as can be verified by use of the similar right triangles.

In figures 2 through 5 it has been tacitly assumed that the distances involved exceeded the unit distance. This is, of course, not required and the constructions proceed in the same manner. For example, if B is less than unity in figure 2, this simply means that the larger triangle becomes the smaller and vice-versa. Similar remarks hold for figures 4 and 5 if A is less than the unit distance.

3. Solution of $\ddot{x} + f(\dot{x})\, g(x) + \phi(\dot{x}) + h(x) = 0$

By suitable applications of the ideas of the first two sections, the trajectory of the equation

$$\ddot{x} + f(\dot{x})\, g(x) + \phi(\dot{x}) + h(x) = 0 \qquad (2)$$

can be constructed. Figure 6 shows the details, and the method is outlined below.

1) Prior to any actual construction, the nonlinear functions $f(\dot{x})$, $g(x)$, $\phi(\dot{x})$, and $h(x)$ are plotted on the axis shown.

2) From the starting point P, the horizontal line FG is drawn. This intersects the $\dot{x}$ axis at D, $f(\dot{x})$ at E and $\phi(\dot{x})$ at F.

3) From the starting point P, the vertical line PC is drawn. This intersects the x axis at A, $h(x)$ at B and $g(x)$ at C.

4) The point B' is laid off such that AB' = AB = $h(x)$.

5) The point F' is laid off such that B'F' = DF = $\phi(x)$.

6) From E the line EH is drawn such that DK = 1.

7) The length AC = $g(x)$ is positioned between the lines EH and EG. (This is shown as A'C'.) Hence EC' = (ED) (AC) = $f(\dot{x})\ g(x)$.

8) Lay off E'F' = EC'. Hence E'A = E'F' + F'B' + B'A = $f(\dot{x})\ g(x) + \phi(\dot{x}) + h(x)$.

9) Draw the line E'P. The slope of this line is PA/E'A = $\dot{x}$/E'A

10) Construct the segment ab normal to E'P at P. The slope of this line is --- E'A/PA = $-\ (f(\dot{x})\ g(x) + \phi(\dot{x}) + h(x))/\dot{x}$ and is therefore the slope ($d\dot{x}/dx = \ddot{x}/\dot{x}$) of the trajectory at P.

The trajectory can be constructed with good accuracy by assuming a new point, (say P') on the line ab close to P. The answer to the question, "How close is close?" is determined by noting whether the new slope at P' is very much different from the slope at P. If it is not, the point P' is "close" to P.

4. Solution of $\dddot{x} + g(\ddot{x}) + \phi(\dot{x}) + h(x) = 0$

To plot the trajectory of a third order system requires a three dimensional phase space. This will be done by plotting the projections of the space trajectory on two planes. These uniquely determine the third projection, hence completely describe the motion of the system. The two projections to be constructed are the $\ddot{x} - \dot{x}$ and $\ddot{x} - x$ curves. The pertinent phase plane equations are

$$\frac{d\ddot{x}}{d\dot{x}} = \frac{\dddot{x}}{\ddot{x}} = -\ \frac{g(\ddot{x}) + \ \phi(\dot{x}) + h(x)}{\ddot{x}} = -\ \frac{N}{\ddot{x}} \tag{3}$$

$$\frac{d\ddot{x}}{dx} = \frac{\dddot{x}}{\dot{x}} = -\ \frac{g(\ddot{x}) + \ \phi(\dot{x}) + h(x)}{\dot{x}} = -\ \frac{N}{\dot{x}}\,. \tag{4}$$

Since both equations have the same numerator N, the amount of construction required for these two equations is considerably less than that required for two unrelated equations. The construction details are shown in figure 7, and the steps are listed below.

1) Join the starting points P,P' and P" by straight lines. PP' intersects the x axis at A and $h(x)$ at B. P'P" intersects the $\ddot{x}$ axis at C and $g(\ddot{x})$ at D.

2) Draw a line PF at P parallel to the x axis. This intersects the $\dot{x}$ axis at E and $\phi(\dot{x})$ at F.

3) Construct the line P"A' parallel to PP'.

4) Lay off $A'B' = AB = h(x)$.

5) Lay off $B'E' = FE = \phi(x)$.

6) Lay off $E'C' = DC = g(\ddot{x})$. Hence $A'C' = A'B' + B'E' + E'C' = h(x) + \phi(\dot{x}) + g(\ddot{x})$.

7) Draw C'P". Since $A'P'' = \ddot{x}$, the slope of $C'P'' = A'P''/A'C' = \ddot{x}/(h(x) + \phi(\dot{x}) + g(\ddot{x})) = \ddot{x}/N$.

8) Construct ab normal to C'P". This line has the slope of the $\ddot{x} - \dot{x}$ trajectory at P".

9) On P'P lay off $P'C'' = AP = \dot{x}$.

10) At C" draw A"C" parallel to the x axis such that $A''C'' = A'C'$.

11) Draw the line A"P'. The slope of this line is $P'C''/A''C'' = \dot{x}/[(h(x) + \phi(\dot{x}) + g(\ddot{x}))] = \dot{x}/N$.

12) Construct de normal to A"P' at P'. This line has the slope of the $\ddot{x} - x$ trajectory at P'.

Note that selecting a new point in any of the planes determines the corresponding points in the other two. This is so since $AP' = A'P'' = \ddot{x}$, $EP = CP' = x$, and $CP'' = AP = \dot{x}$. Hence constructing the trajectories $\ddot{x} - x$ and $\ddot{x} - x$ determines the $\dot{x} - x$ plot.

5. Solution of $x^{(n)} + \sum_{i=1}^{n-1} f_i(x^{(i)}) = 0$

This section extends the results of the previous section to n-dimensional space, ($x^{(k)}$) is the k-th derivative of x, and $f_k(x^{(k)})$ is any function of the k-th derivative). In general, the n-th order system is solved if the (n-1)st and all lower derivatives are known. Hence plots of $x^{(n-1)}$ versus the lower derivatives are sufficient to describe the motion of the system in phase space. Each plot of $x^{(n-1)}/\,x^{(k)}$ has a phase trajectory equation

$$\frac{dx^{(n-1)}}{dx^{k}} = \frac{x^{(n)}}{x^{(k+1)}} = -\frac{N}{x^{(k+1)}} \qquad (5)$$

where $N = \sum_{i=1}^{n-1} f_i(x^{(i)})$. Hence the same numerator can be used to determine all slopes of interest. The suggested method of solution is outlined in figure 8. There it can be seen that n-2 constructions of the type described in the previous section are made. (Each such construction is called a graph.) It should not be concluded that this requires n-2 times the effort. Note that in each graph after the first, one of the required constructions was done in the preceding graph. (Specifically, the plot of $x^{(n-1)}$ versus $x^{(n-k-1)}$, required in the k-th graph, has already been done in the (k-1)th graph.) In addition, the quantity N, which is required for the remaining trajectory, was determined in the first graph. Hence, the construction on all graphs after the first proceeds quite rapidly. The distance N is determined by adding the proper distances from each of the n-2 graphs. Note that the k-th graph $(k \neq 1)$ need contain only the plot of $f_{n-k-2}\,(x^{(n-k-2)})$. That is to say $f_i(x^{(i)})$ is only plotted once.

6. High Order Systems Involving Products

Products of (not necessarily linear) functions often occur in high order systems. By proper use of the methods described in sections 3 and 5, the phase space trajectory of these systems can be constructed. Since the number of possible types of such equations is large, the procedure is best described by means of an example.(4)

Consider Schlichting's equation

$$\dddot{x} + x\,\ddot{x} + (1-\dot{x}^2) = 0 \qquad (6)$$

which is of the form

$$\dddot{x} + f(x)\,g(\ddot{x}) + \phi\,(\dot{x}) = 0 \qquad (7)$$

where $f(x) = x$, $g(\ddot{x}) = \ddot{x}$ and $\phi\,(x) = 1-\dot{x}^2$.

Figure 9 shows the plots necessary to obtain the desired trajectories, and the resulting trajectories. The initial values selected were $\dot{x} = x = 0$ and $\ddot{x} = 1.233$.

7. General Comments

The list of construction steps in sections 2 and 3 makes the construction seem more lengthy and tedious than in actual practice. It should be pointed out that all of the constructions listed need not be performed, but are included only to describe the procedure in a formal manner. The author has found it convenient to measure the pertinent distances by means of marks on an auxiliary piece of paper. In this manner a "running total" can be kept, and has the further feature of providing the distance N at the end of the measurements, thus making available a convenient means of laying off this distance wherever needed.

As far as the higher order cases are concerned, the method has been described assuming a single equation. In practice, one may encounter a number of nonlinear equations of low degrees with some (possibly nonlinear) coupling between them. Solving these equations simultaneously may be impossible or impractical. Let x_1, x_2, ... be the variables in each equation, provided that each equation is of second order (the usual case). The scheme shown in figure 10 will provide the phase space solution. The pertinent phase plane equations can be found by well-known techniques.(5)

8. Summary

By the application of the principles outlined in sections 1 and 2, the feasibility of a graphical solution for a large class of nonlinear systems has been demonstrated in sections 2 through 6. The comments of section 7 indicate that application of these same principles enable solutions of other types of problems not considered in this paper. (Here the ingenuity of the engineer can be used to full advantage.) The method is extremely flexible since the nonlinearities can be given graphically (i.e.: hysteresis, dead-space, back-lash, etc.).

BIBLIOGRAPHY

1. W. H. Pell, "Graphical Solution of Single-Degree-of-Freedom Vibration Problem with Arbitrary Damping and Restoring Forces", Journal of Applied Mechanics, 1957, Vol. 24, pp. 311-312.

2. J. E. Gibson, Nonlinear Automatic Control, McGraw-Hill Book Co., 1963, pp. 244-246.

3. G. C. Reis, "An Extension of Pell's Method", IEEE Trans. of PGAC, July, 1964, p. 315.

4. Y. H. Ku, Analysis and Control of Nonlinear Systems, Ronald Press Co., 1958, pp. 209-211.

5. Y. H. Ku, Ibid., pp. 234-239.

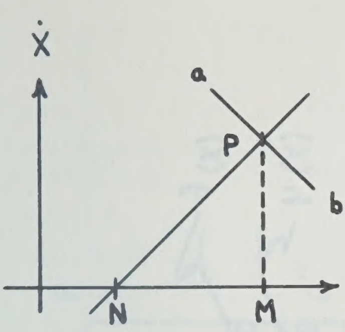


Figure 1

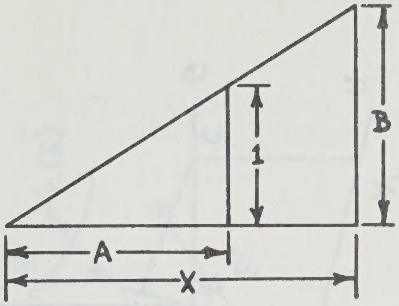


Figure 2

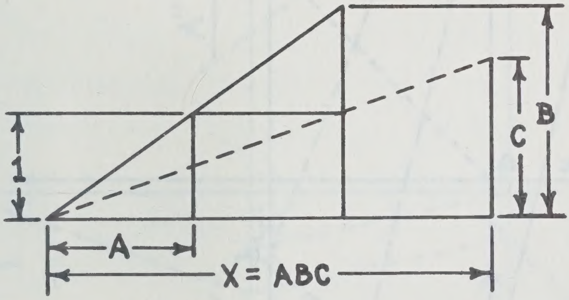


Figure 3

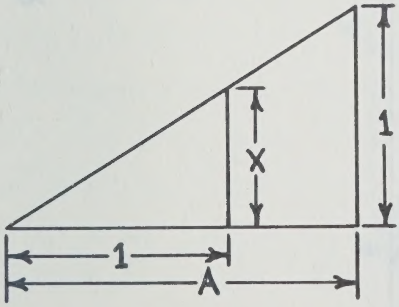


Figure 4

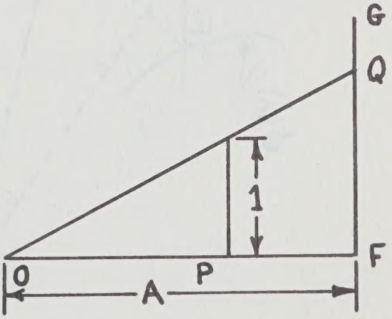


Figure 5

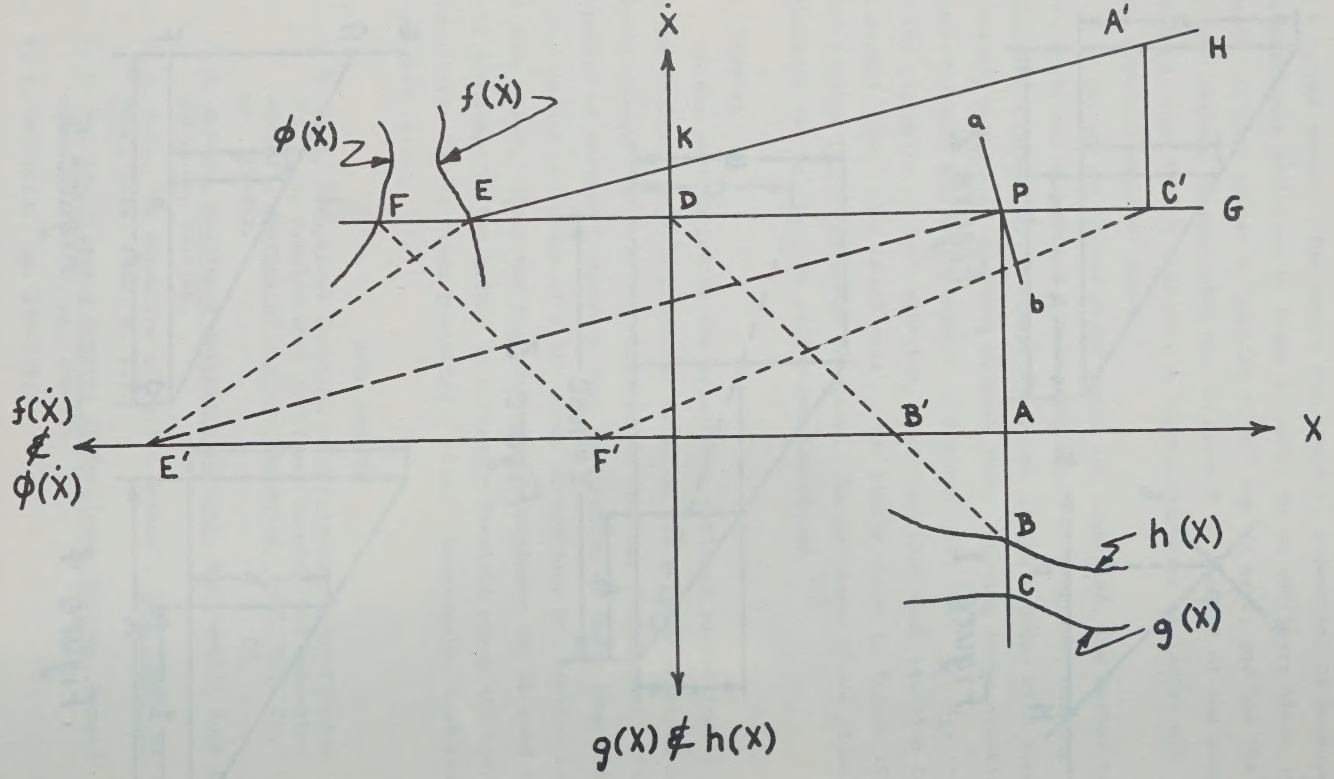

$\dot{X}$
A′
H
$f(\dot{X})$
$\phi(\dot{X})$
K
a
F
E
D
P
C′
G
b
$f(\dot{X}) \notin \phi(\dot{X})$
E′
F′
B′
A
X
B
h (X)
C
g (X)
$g(X) \notin h(X)$

Figure 6

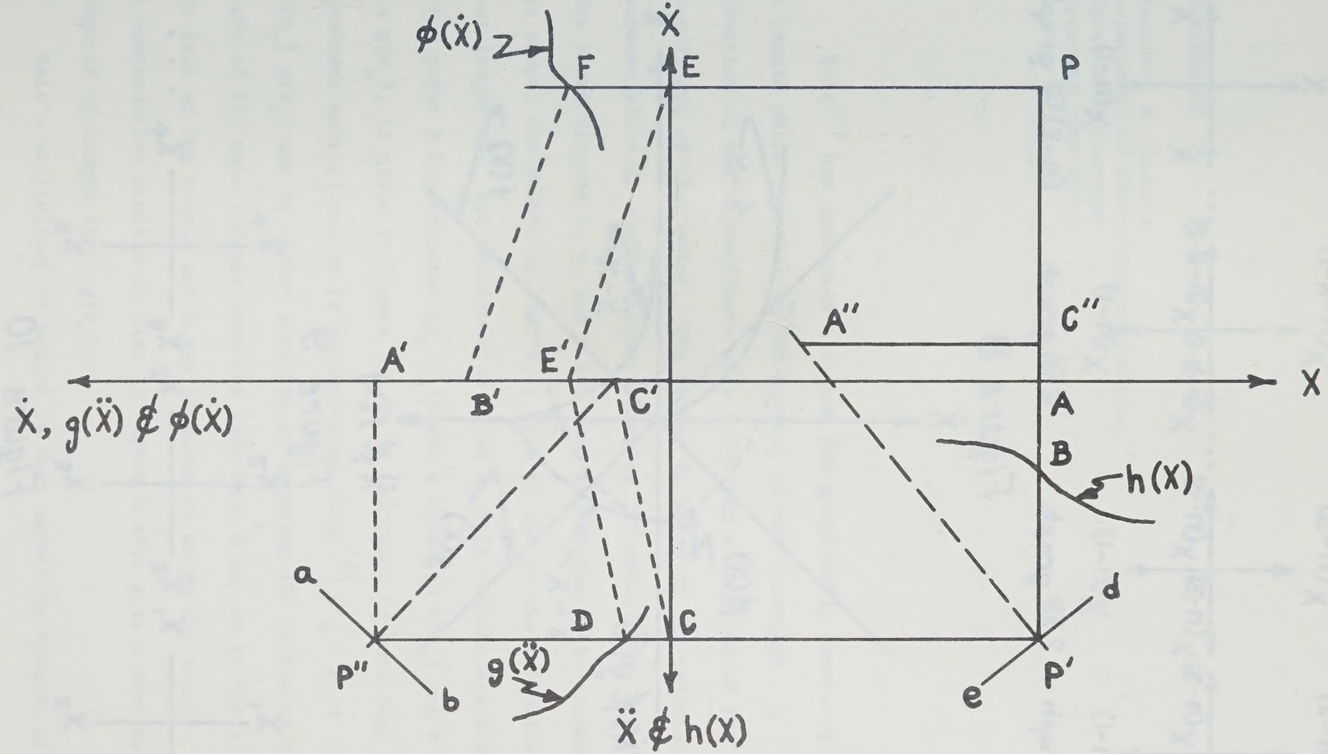
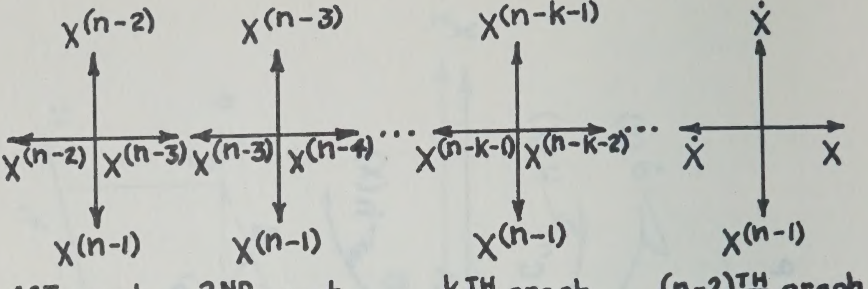

$\phi(\dot{X})$
$\dot{X}$
F
E
P
A''
C''
A'
E'
X
B'
C'
A
$\dot{X}, g(\ddot{X})$ & $\phi(\dot{X})$
B
h(X)
a
d
D
C
P''
$g(\ddot{X})$
b
e
P'
$\ddot{X}$ & h(X)

Figure 7


$X^{(n-2)}$
$X^{(n-2)}$
$X^{(n-3)}$
$X^{(n-1)}$
$X^{(n-3)}$
$X^{(n-3)}$
$X^{(n-4)}$
$X^{(n-1)}$
$X^{(n-k-1)}$
$X^{(n-k-1)}$
$X^{(n-k-2)}$
$X^{(n-1)}$
$\dot{X}$
$\dot{X}$
X
$X^{(n-1)}$

1ST graph 2ND graph kTH graph (n-2)TH graph

Figure 8

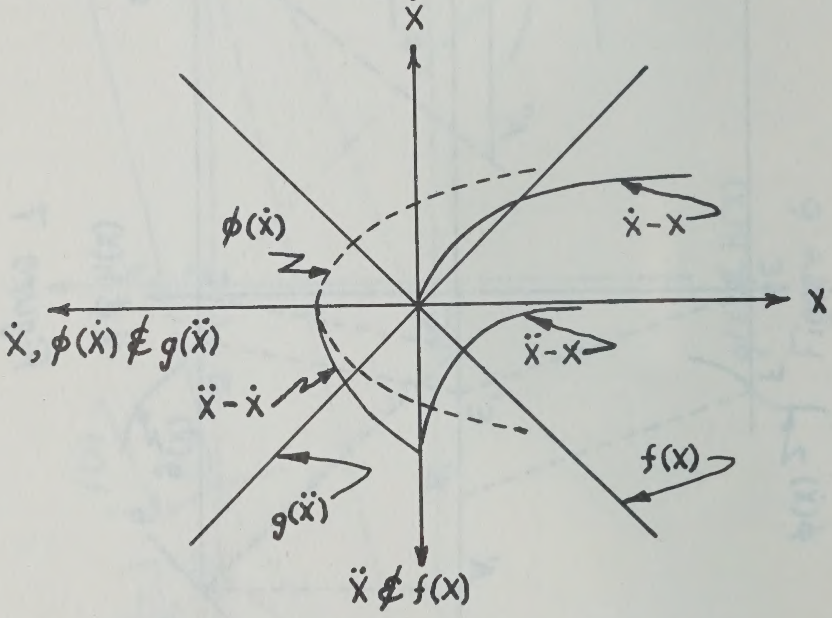

$\dot{X}$
X
$\dot{X}, \phi(\dot{X}) \& g(\ddot{X})$
$\ddot{X} \& f(X)$
$\phi(\dot{X})$
$\dot{X}-X$
$\ddot{X}-X$
$\ddot{X}-\dot{X}$
$f(X)$
$g(\ddot{X})$

Figure 9

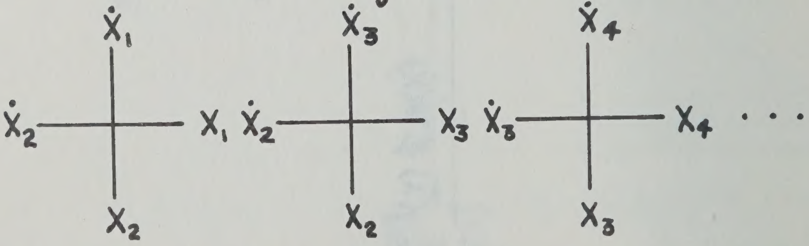

$\dot{X}_1$
$\dot{X}_2$
X_1
X_2
$\dot{X}_3$
$\dot{X}_2$
X_3
X_2
$\dot{X}_4$
$\dot{X}_3$
X_4
X_3
. . .

Figure 10

DISCUSSION

Theodore A. Bickart
Department of Electrical Engineering
Syracuse University
Syracuse, New York

Reis(1) has presented a scheme for graphically constructing a length which is equal to the product of two functions - a function of the x-coordinate of a point and a function of the y-coordinate of that point. The method is based on the relationship between sides of similar right triangles, illustrated in Figure 1. The known lengths 1, a, and b are used to construct the triangles; as shown, the result is that u, a side of one of the triangles, has length ab. The construction given by Reis, which, starting from point P (coordinates (x_o, y_o)) sets a equal to $f(y_o)$ and b equal to $g(x_o)$, is given in Figure 2. It is important to note that four-quadrant multiplication is accomplished since(2) $C'_x - E_x = f(y_o)\ g(x_o)$ may be either positive or negative (C' to the right or left of E); this follows from the fact that E may be to the right or left of the y-axis and C may be above or below the x-axis. A construction similar to that in Figure 2 will give a vertical line segment of length $f(y_o)\ g(x_o)$.

Reis's multiplication method is very appealing because of its simplicity. The variations on the method that are to be presented in this note, hopefully, will add a greater degree of preciseness to the method through elimination of the need to translate $\overline{AC}$ to $\overline{A'C'}$. The first proposed variation locates the horizontal length $f(y_o)\ g(x_o)$ on the x-axis and uses the curve $y - g(x) = 0$ as a construction aid rather than $y + g(x) = 0$; the procedure to be followed is shown in Figure 3. The second variation, which involves no translating of line segments, provides a point having as its x-coordinate the desired product, $f(y_o)\ g(x_o)$, rather than a line segment of length $f(y_o)\ g(x_o)$. This construction is shown in Figure 4.

The illustration in Figure 5 is given so as to make the construction to obtain the product $f(y_o)\ g(x_o)$ stand apart from that portion of the construction, given in Figure 4, which obtains the functions $f(y_o)$ and $g(x_o)$. That is, starting from the point $(f(y_o),\ g(x_o))$ - easily located from the point (x_o,y_o) - the point $(f(y_o)g(x_o),\ 0)$ is located by the steps shown in Figure 5. This multiplication scheme is noticeably simpler than the graphical realization of quarter-square multiplication developed by Bickart.(3,4)

Footnotes

1. G. C. Reis, "On the Construction of Phase-Trajectories", Proceedings of the Second Annual Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, Illinois, September 1964.

2. The subscript x or y, on the capital letter denoting a point, indicates x- or y-coordinate of that point.

3. Theodore A. Bickart, "Nonlinear Analysis in the Phase Plane", Journal of the Franklin Institute, October 1964.

4. Theodore A. Bickart, "Second Order Autonomous, Nonlinear System: A Graphical Solution", Proceedings of the Second Annual Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, Illinois, September 1964, and Syracuse University Department of Electrical Engineering Technical Report TR-64-8, September 1964.

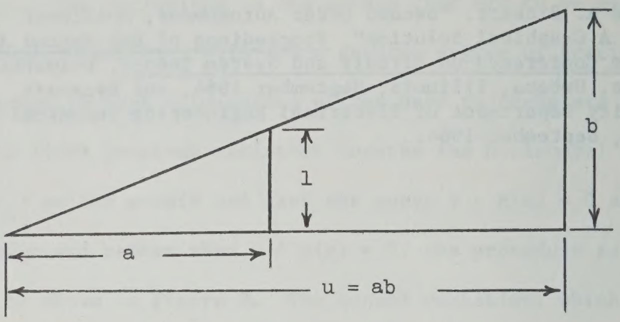
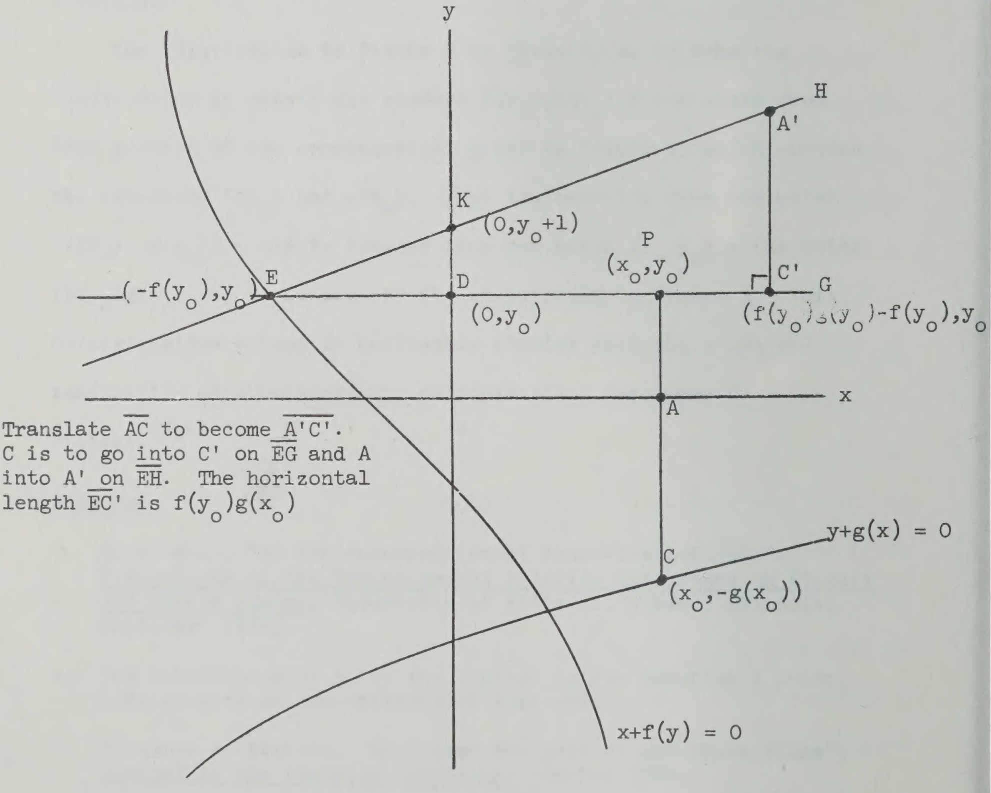


Figure 1 Product Relationship Between Sides of Similar Right Triangles


y
H
A'
K
(0,y_o+1)
P
(x_o,y_o)
C'
G
(-f(y_o),y_o)
E
D
(0,y_o)
(f(y_o)g(y_o)-f(y_o),y_o
x
A
y+g(x) = 0
C
(x_o,-g(x_o))
x+f(y) = 0

Translate $\overline{AC}$ to become $\overline{A'C'}$. C is to go into C' on $\overline{EG}$ and A into A' on $\overline{EH}$. The horizontal length $\overline{EC'}$ is $f(y_o)g(x_o)$

Figure 2 Reis' Method For Finding A Line Segment of Length $f(y_o)g(x_o)$

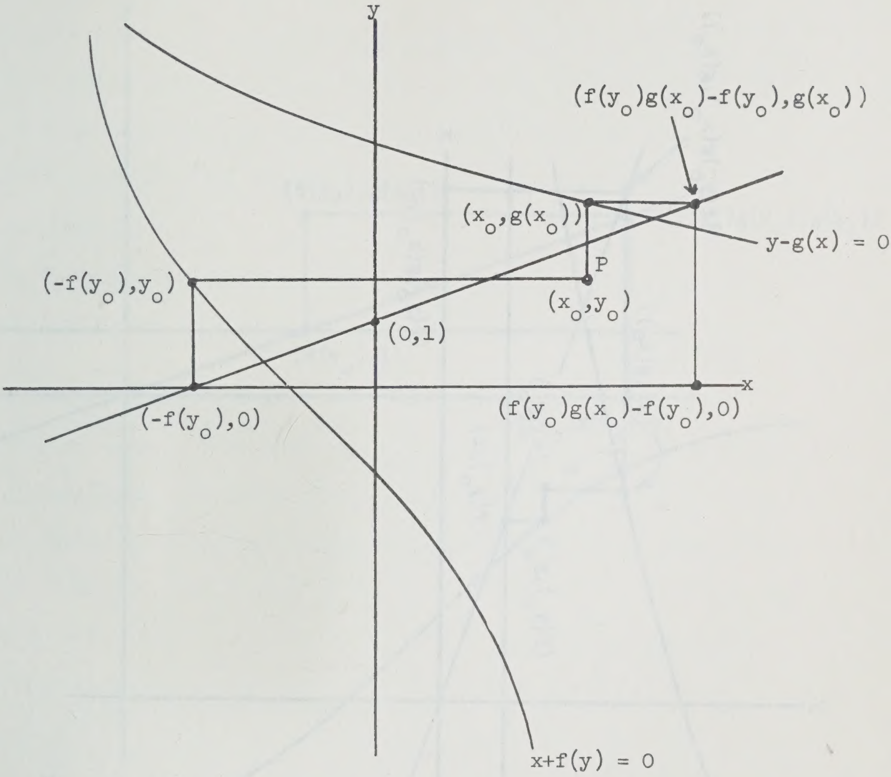

y
x
$(f(y_o)g(x_o)-f(y_o),g(x_o))$
$(x_o,g(x_o))$
$y-g(x) = 0$
P
$(-f(y_o),y_o)$
(x_o,y_o)
$(0,1)$
$(-f(y_o),0)$
$(f(y_o)g(x_o)-f(y_o),0)$
$x+f(y) = 0$

Figure 3 Construction For Finding Line Segment of Length $f(y_o)g(x_o)$ Without The Need To Translate Line Segments

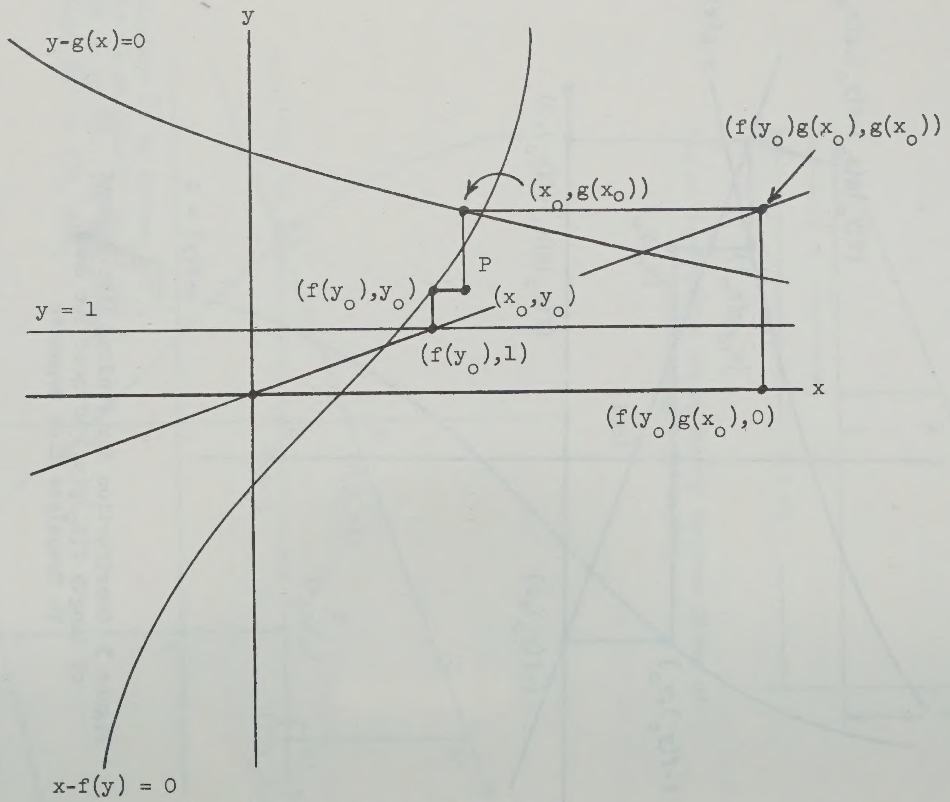

y
y-g(x)=0
$(f(y_o)g(x_o),g(x_o))$
$(x_o,g(x_o))$
P
$(f(y_o),y_o)$
(x_o,y_o)
y = 1
$(f(y_o),1)$
x
$(f(y_o)g(x_o),0)$
x-f(y) = 0

Figure 4 Construction To Locate The Point $(f(y_o)g(x_o),0)$

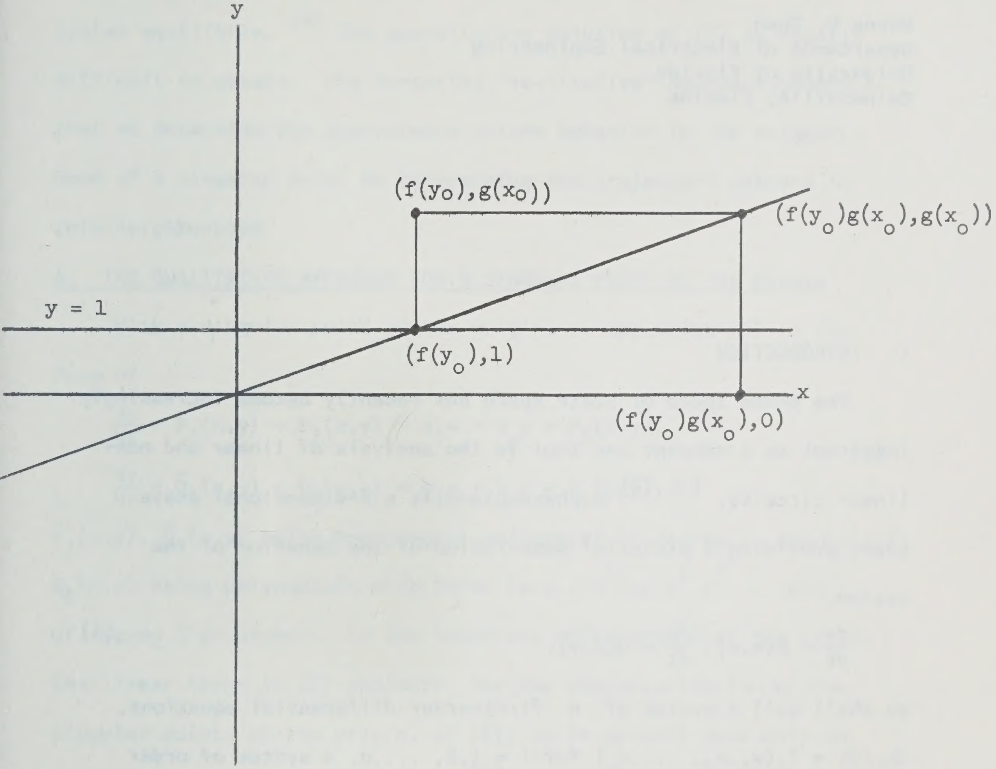

y
(f(yo),g(xo))
(f(yo)g(xo),g(xo))
y = 1
(f(yo),1)
x
(f(yo)g(xo),0)

Figure 5 Construction To Locate $(f(y_o)g(x_o),0)$ Starting From $(f(y_o), g(x_o))$

CHARACTERIZATION OF NON-ELEMENTARY SINGULAR POINTS OF NONLINEAR SYSTEMS IN THE PHASE PLANE

Wayne H. Chen
Department of Electrical Engineering
University of Florida
Gainesville, Florida

I. INTRODUCTION

The phase space or state space has recently become increasingly important as a concept and tool in the analysis of linear and nonlinear circuits.$^{(1),(2)}$ A phase plane is a 2-dimensional phase space providing a pictorial description of the behavior of the system,

$$\frac{dx}{dt} = P(x,y), \; \frac{dy}{dt} = Q(x,y). \tag{1}$$

We shall call a system of n first-order differential equations, $dx_i/dt = f_i(x_1, x_2, \ldots, x_n)$ for $i = 1,2, \ldots, n$, a system of order n; (1) is, therefore, a system of order 2. An ordinary differential equation of order n is, in general, also representable as a system of order n. The solutions of (1) as parametrized curves in the phase plane are called trajectories. A "phase portrait" as a collection of trajectories in a phase region tells us how the system behaves in this region. The phase-plane representation is, therefore, important for the investigation of the behavior of 1) order 2 systems, as well as of 2) order n systems with the phase-space method (3) using phase-plane projections of the space trajectories.

The singular points, S_i, of (1) are the points (x_i, y_i) for which $P(x_i, y_i) = 0$ and $Q(x_i, y_i) = 0$, and are the positions of system equilibria. (4) The quantitative solution of (1) is usually difficult to obtain. The so-called "qualitative" approach suggests that we determine the approximate system behavior in the neighborhood of a singular point by determining the trajectory pattern in this neighborhood.

A. THE QUALITATIVE APPROACH FOR A SINGULAR POINT AT THE ORIGIN

With a singular point at the origin, we may write (1) in the form of

$$\frac{dx}{dt} = \bar{P}_1(x,y) + P_2(x,y) = a_1x + a_oy + P_2(x,y),$$
$$\frac{dy}{dt} = \bar{Q}_1(x,y) + Q_2(x,y) = b_1x + b_oy + Q_2(x,y), \tag{2}$$

$\bar{P}_1(x,y)$, $\bar{Q}_1(x,y)$ being homogeneous polynomial of degree 1, and $P_2(x,y)$ $Q_2(x,y)$ being polynomials with terms (e.g., $x^2, xy, y^2, x^3, \ldots$, etc.) of degree 2 or higher. In the immediate neighborhood of the origin the linear terms in (2) dominate; for the characterization of the singular point, at the origin, of (2), we in general need only to consider its linear approximation (4)

$$\frac{dx}{dt} = \bar{P}_1(x,y) = a_1x + a_oy\ , \quad \frac{dy}{dt} = \bar{Q}_1(x,y) = b_1x + b_oy. \tag{3}$$

The reader's familiarity with the singular points, at the origin, of (2) and (3), as discussed in standard nonlinear textbooks, (1) (5) (6) and summarized in a tutorial paper by Minorsky, (4) is assumed.

For determining the trajectory pattern in the neighborhood of the singular point, at the origin, of (2) or (3), we need only to follow this simple procedure:

1) Obtain the characteristic roots s_1, s_2 from the characteristic equation, $s^2-(a_1+b_o)\,s+(a_1b_o-a_ob_1)=0$, and interpret them according to fixed rules$^{(4)}$ in determining the type of the singular point. Figure 1 depicts the types of these singular points.

2) Obtain the real linear factors, $y-\alpha_i x=0$ or $y=\alpha_i x$, $i=1,2,\ldots,h$, of this auxiliary equation$^{(7)}$

$$x\bar{Q}_n(x,y) - y\bar{P}_n(x,y) = 0 \tag{4}$$

for n=1 (and $h\leq 2$); equation (4) is also intended for later use for any systems with $n\geq 1$. These real linear factors represent the "asymptotes" of (2) or (3), namely, 1) the asymptotic trajectories, e.g., AOB and COD in Figures 1(a) through (c), for (3); 2) the asymptotic tangents, e.g., A'OB' and C'OD' in Figures 1(g) and (h), for (2). The asymptotic trajectories for (2) are curves, e.g., AOB and COD in Figures 1(h) and (g), and are approximately represented by the asymptotic tangents in the immediate neighborhood of the origin.

3) Sketch an approximate unparametrized (i.e., arrowheadless) trajectory pattern according to the type of the singular point and the positions of its asymptotes (if any) as determined above.

4) Parametrize an arbitrary trajectory T or an asymptote A_i: $y=\alpha_i x$ by first 1) choosing a point R at (x_o,y_o) on T or A_i, close to the origin, and for $x_o>0$; 2) computing $\bar{P}_n(x_o,y_o)$ or $\bar{P}_n(1,\alpha_i)$ 3) drawing an arrowhead along T or A_i in the direction of increasing (decreasing) x as indication of parametrization for a positive (negative) $\bar{P}_n(x_o, y_o)$

or $\bar{P}_n(1, \alpha_i)$. This step results from the definition $dx/dt\Big|_R = \lim_{\Delta t \to o} \Delta x/\Delta t = \bar{P}_n(x_o,y_o) = x_o^{\ n}\, \bar{P}_n(1,\alpha_i)$ as the rate of change of x with respect to t; for a positive (negative) $dx/dt\big|_R$, x increases (decreases) as t increases.

5) Parametrize all trajectories, neighboring to the already-parametrized trajectory T or A_i, in the same general direction; extend parametrization in the same manner to all trajectories in the neighborhood of the singular point concerned.

The result obtained above with the qualitative approach, if desired, may be refined with, or used as reference for, the graphical method of isoclines (8)

B. THE QUALITATIVE APPROACH FOR A SINGULAR POINT NOT AT THE ORIGIN

If the singular point of (1) is situated at (x_i,y_i) other than the origin, we may introduce a translation $x = \bar{x} + x_i$, $y = \bar{y} + y_i$ such that the singular point is at the origin of the new $(\bar{x},\bar{y})$ - coordinate system. We may now repeat the procedure in the preceding paragraphs to determine the trajectory pattern in the neighborhood of the singular point.

C. THE QUALITATIVE APPROACH FOR SEVERAL SINGULAR POINTS

To investigate the system behavior in a phase region containing several singular points, we may use the techniques described above to obtain trajectory patterns in the neighborhoods of the various singular points and then interconnect these patterns into a single phase portrait. The method of isoclines may be used, when necessary, to help making the interconnections and refining the portrait. Figure 2 depicts a phase portrait in a region containing a stable focus F, a saddle point S and a stable node N.

D. SYSTEMS WITHOUT LINEAR TERMS

To investigate the behavior of the system

$$\frac{dx}{dt} = -6x + 8y + 3x^2 - 7y^2 + 2y^3,$$
$$\frac{dy}{dt} = 6x - 2y - 7x^2 + 4xy - y^2 + 2x^3, \qquad (5)$$

in a phase region containing its singular points at (0,0) and (1,1), we first 1) determine, with the procedure in Section I-A above, the trajectory pattern (which will be given in Figure 7) in the neighborhood of the singular point at (0,0); then 2) follow the procedure in Section I-B by introducing a translation $x = \bar{x} + 1$, $y = \bar{y} + 1$ and modifying (5) into

$$\frac{d\bar{x}}{dt} = 3\bar{x}^2 - \bar{y}^2 + 2\bar{y}^3,$$
$$\frac{d\bar{y}}{dt} = -\bar{x}^2 + 4\bar{x}\bar{y} - \bar{y}^2 + 2\bar{x}^3 . \qquad (6)$$

Since (6) has no linear terms and cannot be studied with the procedure in Section I-A, it needs special attention and will be treated in Section II.

II. CHARACTERIZATION OF NON-ELEMENTARY SINGULAR POINTS

We shall now consider

$$\frac{dx}{dt} = \bar{P}_n(x,y) + P_{n+1}(x,y) = a_n x^n + \ldots + a_1 x y^{n-1} + a_o y^n + P_{n+1}(x,y),$$
$$\frac{dy}{dt} = \bar{Q}_n(x,y) + Q_{n+1}(x,y) = b_n x^n + \ldots + b_1 x y^{n-1} + b_o y^n + Q_{n+1}(x,y). \qquad (7)$$

As in the previous case of (2) and (3), the lowest-degree terms in (7) dominate in the immediate neighborhood of the origin; for the characterization of the singular point at the origin of (7), we may consider its lowest-degree approximation

$$\frac{dx}{dt} = \bar{P}_n(x,y) = a_n x^n + a_{n-1} x^{n-1} y + \ldots + a_1 x y^{n-1} + a_o y^n ,$$
$$\frac{dy}{dt} = \bar{Q}_n(x,y) = b_n x^n + b_{n-1} x^{n-1} y + \ldots + b_1 x y^{n-1} + b_o y^n ; \qquad (8)$$

A simple procedure for characterization will be introduced in the following paragraphs and then justified in Section IV.

A. STEP 1: DETERMINE THE POSITION OF ASYMPTOTES

We first 1) obtain from (4) its real linear factors

$$y - \alpha_i x = o \quad \text{or} \quad y = \alpha_i x, \quad i = 1,2,\ldots,h, \qquad (9)$$

as well as its complex factors $y - \alpha_i x = o$, $i = (h+1),(h+2),\ldots,(n+1)$, and then 2) draw the asymptotes of (7) or (8) with (9) as straight lines passing through the origin and approximately dividing its neighborhood into 2h regions.

B. STEP 2: OBTAIN THE CHARACTERISTIC RATIOS

We now compute the characteristic ratio

$$\lambda_i = s_1^{(i)} / s_2^{(i)} , \qquad i = 1,2,\ldots,h, \qquad (10)$$

associated with each asymptote $y = \alpha_i x$, where 1) for $a_o \neq o$ in (7) or (8),

$$s_1^{(i)} = a_o \, [(\alpha_i - \alpha_1)(\alpha_i - \alpha_2) \ldots (\alpha_i - \alpha_{i-1})(\alpha_i - \alpha_{i+1}) \ldots (\alpha_i - \alpha_{n+1})] / (1 + \alpha_i^2)^{(n-1)/2} ,$$
$$s_2^{(i)} = -\bar{P}_n(1,\alpha_i)/(1+\alpha_i^2)^{(n-1)/2} , \qquad (11)$$

for $i = 1,2,\ldots,h, \bar{P}_n(1,\alpha_i)$ being the expression $\bar{P}_n(x,y)$ with $x = 1$ and $y = \alpha_i$, and 2) for $a_o = o$ and designating $\alpha_1 = \infty$ (because the associated asymptote $y = \alpha_1 x$ is actually $x = y/\alpha_1 = o$),

$$s_1^{(i)} = (a_1 - b_o) \, [(\alpha_i - \alpha_2)(\alpha_i - \alpha_3) \ldots (\alpha_i - \alpha_{n-1})(\alpha_i - \alpha_{i+1})(\alpha_i - \alpha_{n+1})] / (1+\alpha_i^2)^{(n-1)/2} \qquad (12)$$

and $s_2^{(i)}$ as in (11) for $i = 2,3, \ldots, h$, while

$$s_1^{(1)} = b_o - a_o, \quad s_2^{(1)} = -b_o \quad (\text{i.e., } i = 1). \tag{13}$$

C. STEP 3: SKETCH AND PARAMETRIZE THE TRAJECTORY PATTERN

Sketch an unparametrized trajectory pattern in each region R_{jk} approximately bounded by two neighboring asymptotes A_j and A_k, which are respectively associated with characteristic ratios λ_j and λ_k, according to the following rules:

1) For $\lambda_j > 0$ (or $\lambda_k > 0$), the unparametrized (i.e., arrowheadless) trajectories in R_{jk} approach the origin in the direction of A_j (or A_k).

2) For $\lambda_j < 0$ (or $\lambda_k < 0$), the unparametrized trajectories in R_{jk} recede from the origin in the general direction of A_j (or A_k).

Illustrations are provided for (8) in Figures 3(a), (c) and (e), and for (7) in Figures 3(b), (d) and (f). These illustrations suggest three different types of regions, i.e., 1) an open nodal region for $\lambda_j > 0$ and $\lambda_k < 0$ (or $\lambda_j < 0$ and $\lambda_k > 0$), 2) a closed nodal region for $\lambda_j > 0$ and $\lambda_k > 0$, and 3) a saddle region for $\lambda_j < 0$ and $\lambda_k < 0$.

With all 2h regions sketched, we now have an unparametrized trajectory pattern in the neighborhood of the singular point at the origin, which may be parametrized with the procedure described in paragraphs 4) and 5) in Section I-A.

III. ILLUSTRATIONS AND REMARKS

A. ILLUSTRATION 1

For

$$\frac{dx}{dt} = \bar{P}_2(x,y)+P_3(x,y) = x^2 + y^2 + P_3(x,y), \; P_3(x,y) \equiv 0,$$
$$\frac{dy}{dt} = \bar{Q}_2(x,y)+ Q_3(x,y)= 2xy + Q_3(x,y), \; Q_3(x,y) \equiv 0, \tag{14}$$

we may readily use the new procedure in Section II, 1) obtaining its asymptotes $y = \alpha_i x$ for $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = -1$; 2) determining its characteristic ratios $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = -1$; 3) sketch its unparametrized trajectory pattern in Figure 4(a); 4) parametrize the asymptote A_1 by computing $\bar{P}_2(1,\alpha_1) = 1 > 0$ in accordance with the procedure in Section I-A and extending parametrization to other trajectories, as indicated in Figure 4(a). For $P_3(x,y) \not\equiv 0$ and $Q_3(x,y) \not\equiv 0$, the asymptotes A_1, A_2 and A_3 are asymptotic tangents rather than asymptotic trajectories.

B. SOME REMARKS

In the preceding illustration we used seven characteristic numbers, $[\alpha_1, \alpha_2, \alpha_3; \lambda_1, \lambda_2, \lambda_3; \bar{P}_2(1,\alpha_1)] = [0,1,-1; (+),(-),(-); (+)]$, which collectively characterize the singular point, at the origin, for an order 2 system. In general, we may use (2h+1) characteristic numbers:

$$[\, \alpha_1, \alpha_2, \ldots, \alpha_h; \lambda_1, \lambda_2, \ldots, \lambda_h; \bar{P}_n(1, \alpha_i) \,] \tag{15}$$

to characterize the singular point of (7) or (8) at the origin, where 1) $h = n+1$ if (4) has no complex linear factors, 2) $h = n-1$ if one pair of complex factors, 3) $h = n-3$ if two pairs, ..., etc. For n odd, there is a special case that (4) has no real linear factors and the singular point is a focus or center.

C. ADDITIONAL ILLUSTRATIONS

Figures 4 and 5 depict some non-elementary singular points for the system in (8) with $n = 2$ and $n = 3$ respectively.

D. ILLUSTRATION OF USING THE NEW PROCEDURE FOR A LINEAR SYSTEM

For a linear system

$$\frac{dx}{dt} = \bar{P}_1(x,y) = -x-2y, \quad \frac{dy}{dt} = \bar{Q}_1(x,y) = -4x + y \quad , \tag{16}$$

we may also use the new procedure in Section II, obtaining its characteristic numbers, $[\alpha_1, \alpha_2; \lambda_1, \lambda_2; \bar{P}_1(1, \alpha_1)] = [1, -2; (-), (-); (-)]$ and then its phase portrait in Figure 6. By using the earlier procedure in Section I-A, we may readily verify this result. The procedure in Section II is, therefore, also applicable to the system in (2) or (3).

E. ILLUSTRATION 2

For the system in (5) remarked about in Section I-D, we now obtain its phase portrait in Figure 7.

IV. THE TRANSFORMATIONS THAT ARE RESPONSIBLE FOR THE NEW PROCEDURE

We shall now introduce the transformations $^{(9)}$ that are responsible for, and therefore justify, the new procedure of characterization in Section II.

A. THE Π_1 - χ_4 TRANSFORMATION

Let us first write the unparametrized version of (8),

$$\bar{Q}_n(x,y)dx - \bar{P}_n(x,y)dy = 0, \tag{17}$$

defined in the Π_1- or (x,y) - plane. We shall define a transformation M that 1) maps the Π_1-plane unto the surface of a unit sphere χ_4 and the origin of Π_1 into the equator $\bar{0}$ of χ_4, and 2) transform (17) into a new system whose trajectories are described on χ_4.

Transformation $M = T_3T_2T_1$ may be considered to consist of three projective transformations T_1, T_2 and T_3 carried out in succession, i.e., 1) T_1 as indicated in Figure 8(a), and described by $x/1 = \tilde{x}/\tilde{z}$, $y/1 = \tilde{y}/\tilde{z}$ for $\tilde{z} \geq o$ and $\tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2 = 1$; 2) T_2 and T_3, as indicated in Figures 7(b) and (c) and collectively described by $\tilde{x} = 2\bar{x}\bar{z}$, $\tilde{y} = 2\bar{y}\bar{z}$, $\tilde{z} = 1-2\bar{z}^2$ for $1/\sqrt{2} \geq \bar{z} \geq -1/\sqrt{2}$ and $\bar{x}^2 + \bar{y}^2 + \bar{z}^2 = 1$.

Transformation M transforms (17) into this system on $\mathcal{X}_4$:

$$\begin{aligned} &A(\bar{x},\bar{y},\bar{z})\,d\bar{x} + B(\bar{x},\bar{y},\bar{z})\,d\bar{y} + C(\bar{x},\bar{y},\bar{z})\,d\bar{z} = 0, \\ &\bar{x}^2 + \bar{y}^2 + \bar{z}^2 = 1, \end{aligned} \qquad (18)$$

where $A(\bar{x},\bar{y},\bar{z}) = \bar{z}\,[\,(1-2\bar{z}^2)\bar{Q}_n(\bar{x},\bar{y})\,]$, $B(\bar{x},\bar{y},\bar{z}) = -\bar{z}\,[\,(1-2\bar{z}^2)\bar{P}_n(\bar{x},\bar{y})\,]$, $C(\bar{x},\bar{y},\bar{z}) = (1+2\bar{z}^2)\,[\,\bar{x}\bar{Q}_n(\bar{x},\bar{y}) - \bar{y}\bar{P}_n(\bar{x},\bar{y})\,]$, and $\bar{P}_n(\bar{x},\bar{y})$ and $\bar{Q}_n(\bar{x},\bar{y})$ are $\bar{P}_n(x,y)$ and $\bar{Q}_n(x,y)$ in (8) and (17) with x,y replaced by $\bar{x},\bar{y}$.

B. INVESTIGATION OF ELEMENTARY SINGULAR POINTS ON THE EQUATOR $\bar{O}$ OF $\mathcal{X}_4$

The singular points of the system in (18) on $\mathcal{X}_4$ must satisfy $A(\bar{x},\bar{y},\bar{z}) = 0$, $B(\bar{x},\bar{y},\bar{z}) = 0$, $C(\bar{x},\bar{y},\bar{z}) = 0$. For the singular points on the equator $\bar{O}$ of $\mathcal{X}_4$ in Figure 8 (c), we further impose $\bar{z} = 0$, obtaining

$$\bar{x}\bar{Q}_n(\bar{x},\bar{y}) - \bar{y}\bar{P}_n(\bar{x},\bar{y}) = 0. \qquad (19)$$

By factoring (19) into $K\,[(\bar{y}-\alpha_1\bar{x})(\bar{y}-\alpha_2\bar{x})\ldots(\bar{y}-\alpha_h\bar{x})]\times$ [complex factors], the positions of these singular points satisfy

$$\bar{y} = \alpha_i\bar{x}, \quad \bar{z} = o, \qquad i = 1,2,\ldots,h, \qquad (20)$$

as demonstrated in Figure 9.

To investigate each of these singular points, say, S_i, in Figure 9, we first 1) rotate the $\bar{x}$- and $\bar{y}$-axes to make S_i the $(\dot{x},\dot{y},\dot{z}) = (1,0,0)$ point in the new coordinate system with the transformation T_a: $\bar{x} = \dot{x}\cos\theta_i - \dot{y}\sin\theta_i = (\dot{x} - \alpha_i\dot{y})/\sqrt{1+\alpha_i^2}$, $\bar{y} = \dot{x}\sin\theta_i + \dot{y}\cos\theta_i = (\alpha_i\dot{x} + \dot{y})/\sqrt{1+\alpha_i^2}$, $\bar{z} = \dot{z}$; 2) project the representation on $\mathcal{X}_4$ in the neighborhood of S_i into a Π_2 -

plane, which is tangential to $\mathcal{X}_4$ at S_i, with the transformation T_b: $\dot{x} = 1, \dot{y} = \hat{y}$, $\dot{z} = \hat{z}$, obtaining this system in $\prod_2$:

$$\hat{Q}(\hat{y},\hat{z})d\hat{y} - \hat{P}(\hat{y},\hat{z})d\hat{z} = 0; \tag{21}$$

then 3) introduce τ as the time parameter, obtaining a parametrized system in $\prod_2$, $d\hat{y}/d\tau = \hat{P}(\hat{y},\hat{z})$, $d\hat{z}/d\tau = \hat{Q}(\hat{y},\hat{z})$, which is then approximated into

$$\frac{d\hat{y}}{d\tau} = s_1^{(i)}\hat{y} + \left[\begin{matrix}\text{non-linear}\\ \text{terms}\end{matrix}\right], \quad \frac{d\hat{z}}{d\tau} = s_2^{(i)}\hat{z} + \left[\begin{matrix}\text{non-linear}\\ \text{terms}\end{matrix}\right], \tag{22}$$

where $s_1^{(i)}$ and $s_2^{(i)}$ are the expressions in (11) or (12) and (13).

The linear approximation of (22) is the canonical form (1) of a linear system like (3): $s_1^{(i)}$ and $s_2^{(i)}$ are its characteristic roots. Since $s_1^{(i)}$ and $s_2^{(i)}$ are real quantities, the singular point S_i at the origin of the $\prod_2$-plane is a saddle point or node (4) with the $\hat{y}$- and $\hat{z}$-axes as its asymptotes as depicted in Figure 9.

The singular point S_i on the equator $\bar{0}$ of $\mathcal{X}_4$, whose projection in $\prod_2$ is the above described singular point, must also be a saddle point or node with the equator $\bar{0}$ (which maps into the $\hat{y}$-axis of $\prod_2$) and $\bar{y} = \alpha_i\bar{x}$ (which maps into the $\hat{z}$-axis of $\prod_2$) as asymptotes.

C. JUSTIFICATION OF THE NEW PROCEDURE IN SECTION II FOR CHARACTERIZING SINGULAR POINTS

Having investigated all singular points S_i on the equator $\bar{0}$ of $\mathcal{X}_4$ as described above, we obtain a typical phase representation as depicted in Figure 10(a). By introducing a projective transformation T_4 that maps the equator $\bar{0}$ of $\mathcal{X}_4$ into a circle 0 of the $\prod_5$-plane as shown in Figure 10(b), we obtain the phase representation in Figure 10(c) with saddle points and nodes on the circle 0 of $\prod_5$.

The equator $\bar{0}$ of χ_4 and the circle 0 of Π_5 in Figures 10(a) and (c) are the maps of the origin 0 of Π_1 in Figures 8(a) and 10(d) under projective transformations. Intuitively, the phase portrait in Π_1 could be obtained by shrinking the circle 0 of Π_5 in Figure 10(c) into a point and this phase portrait is displayed in Figure 10(d).

Let us compare Figures 10(a) and (d) for the invariant properties of the transformations. The asymptotes $\bar{y} = \alpha_1\bar{x}$ and $\bar{y} = \alpha_3\bar{x}$ associated with saddle points indicate the directions of the unparametrized trajectories (i.e., trajectories less the arrowheads) receding from these saddle points on the equator $\bar{0}$ of χ_4; the asymptotes $y = \alpha_1 x$ and $y = \alpha_3 x$ indicate the directions of the unparametrized trajectories receding from the origin 0 of Π_1, which is the inverse map of $\bar{0}$ of χ_4. The asymptote $\bar{y} = \alpha_2\bar{x}$ associated with nodes indicates the direction of the unparametrized trajectories approaching these nodes on the equator $\bar{0}$ of χ_4; the asymptote $y = \alpha_2 x$ indicates the direction of the unparametrized trajectories approaching the origin 0 of Π_1, which is the inverse map of $\bar{0}$ of χ_4.

Because of 1) the correspondence between the asymptotes $\bar{y} = \alpha_i\bar{x}$ on χ_4 and $y = \alpha_i x$ in Π_1 and 2) the relation in (19), Step 1 of the new procedure in Section II is established. Because the nature of the asymptote $y = \alpha_i x$ in Π_1 is determined by the type of the singular point S_i (associated with $\bar{y} = \alpha_i\bar{x}$), which is in turn determined by 1) its characteristic roots $s_1^{(i)}$ and $s_2^{(i)}$ in (22), (11), or (12) and (13) or 2) its ratio $\lambda_i = s_1^{(i)}/ s_2^{(i)}$ in (10) being positive (for a node) or negative (for a saddle point), Steps

2 and 3 in Section II are justified.

REFERENCES

1. Y. H. Ku, Analysis and Control of Nonlinear Systems. The Ronald Press Co., New York, N.Y.; 1958.
2. L. A. Zadeh and C. A. Desoer, Linear System Theory: The State Space Approach. McGraw-Hill Book Co., Inc., New York, N. Y.; 1963.
3. Y. H. Ku, "A Method of Solving Third and Higher Order Nonlinear Differential Equations", Journal of Franklin Institute, vol. 256, pp. 229 - 244; September 1953.
4. N. Minorsky, 'Theoretical Aspects of Nonlinear Oscillations", IRE Trans. on Circuit Theory, vol. CT-7, pp. 368 - 381; December 1960.
5. N. Minorsky, Nonlinear Oscillations. D. Van Nostrand Co., Inc., Princeton, N. J.; 1962.
6. W. J. Cunningham, Introduction to Nonlinear Analysis. McGraw-Hill Book Co., Inc., New York, N. Y.; 1958.
7. I. Bendixson, "Sur Les Courbes Defines Par Des Equations Differentielles", Acta Mathematica, Bd. 24, pp. 1-88; 1901.
8. B. Van der Pol, "On Relaxation Oscillations", Philosophical Magazine, 7th Series, Vol. 2; November 1926.
9. W. H. Chen, On the Singular Points of Homogeneous Systems of Ordinary Differential Equations, Ph.D. thesis (mathematics), University of Washington; August 1952.

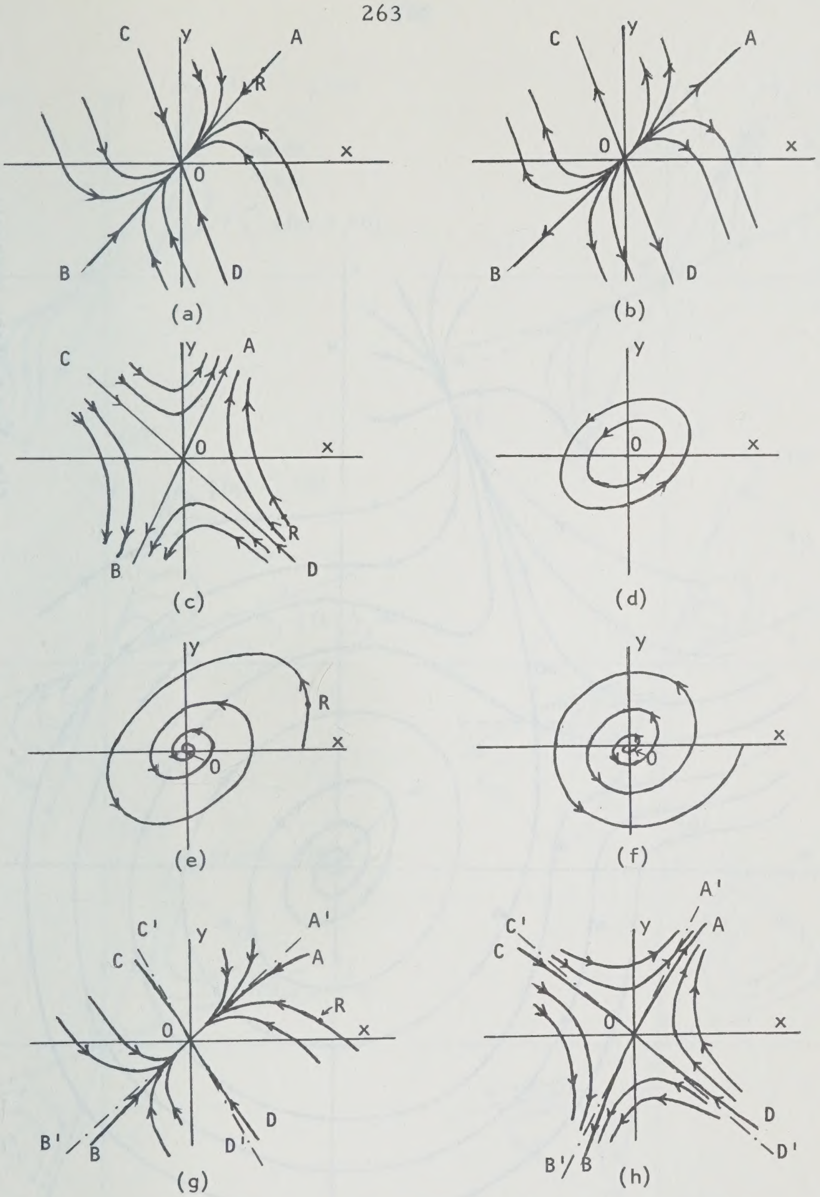

C
y
A
R
x
0
B
D
(a)
C
y
A
0
x
B
D
(b)
C
y
A
0
x
R
B
D
(c)
y
0
x
(d)
y
R
x
0
(e)
y
x
0
(f)
C'
y
A'
C
A
R
0
x
B'
B
D
D'
(g)
A'
C'
y
A
C
0
x
D
D'
B'
B
(h)

Fig. 1 - Elementary singular points at the origin. (a) Stable node. (b) Unstable node. (c) Saddle point. (d) Center. (e) Stable focus. (f) Unstable focus. (g) Another stable node. (h) Another saddle point.

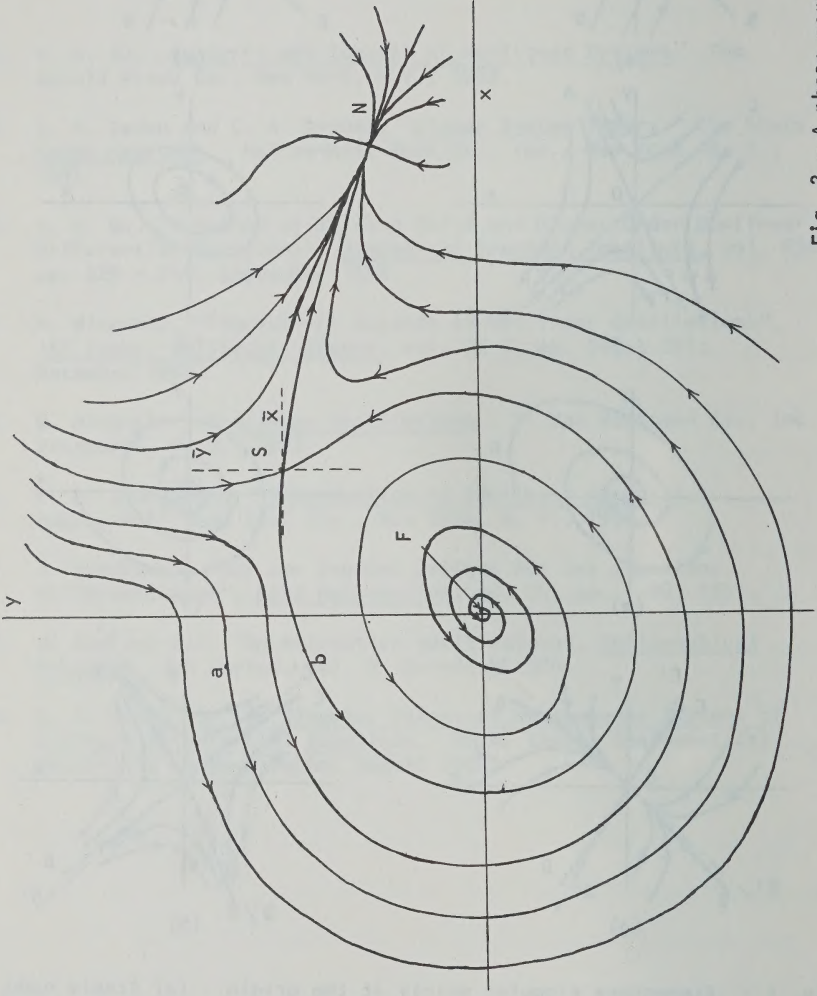

y
x
N
S
$\bar{y}$
$\bar{x}$
F
a
b

Fig. 2 - A phase portrait.

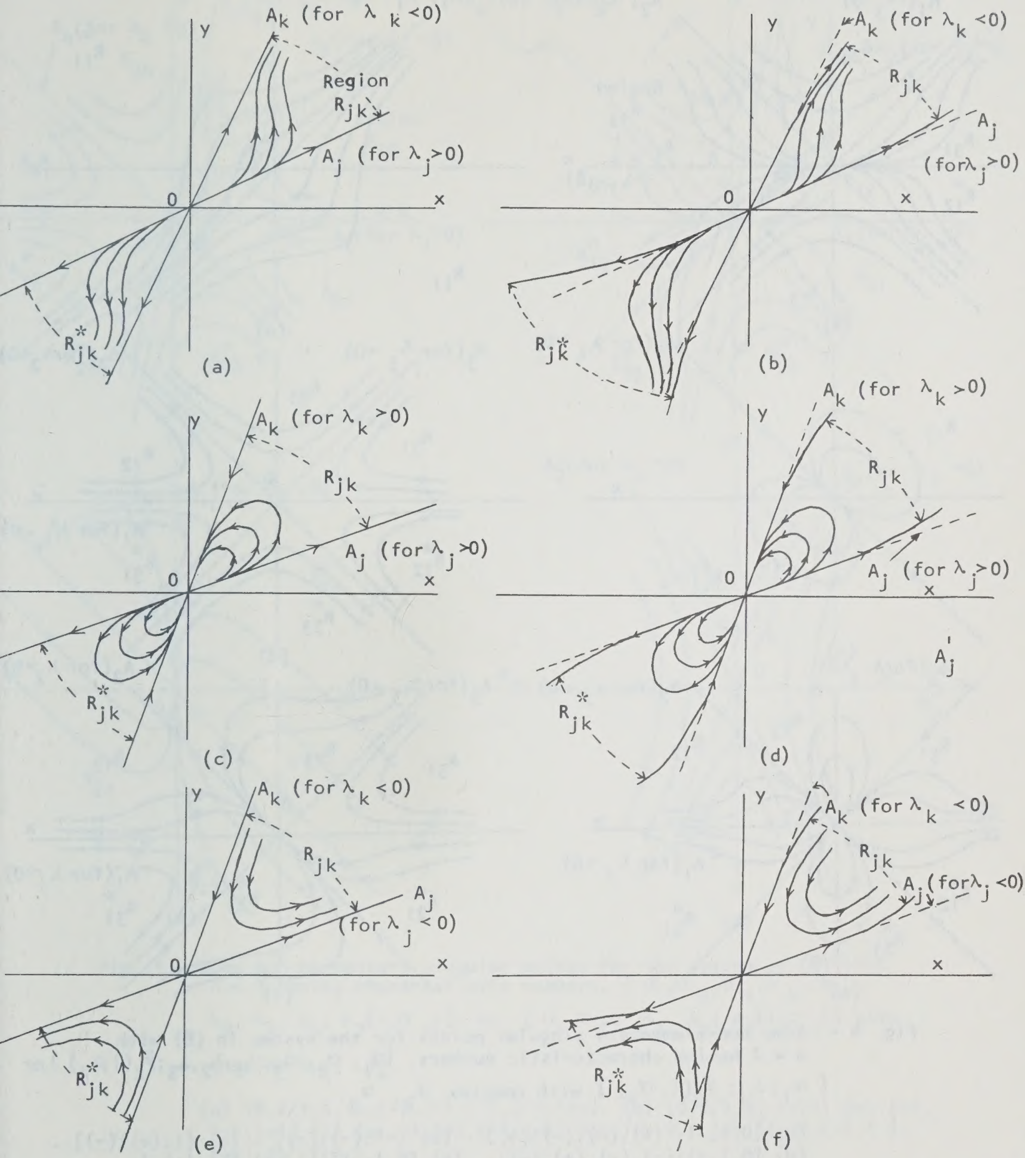
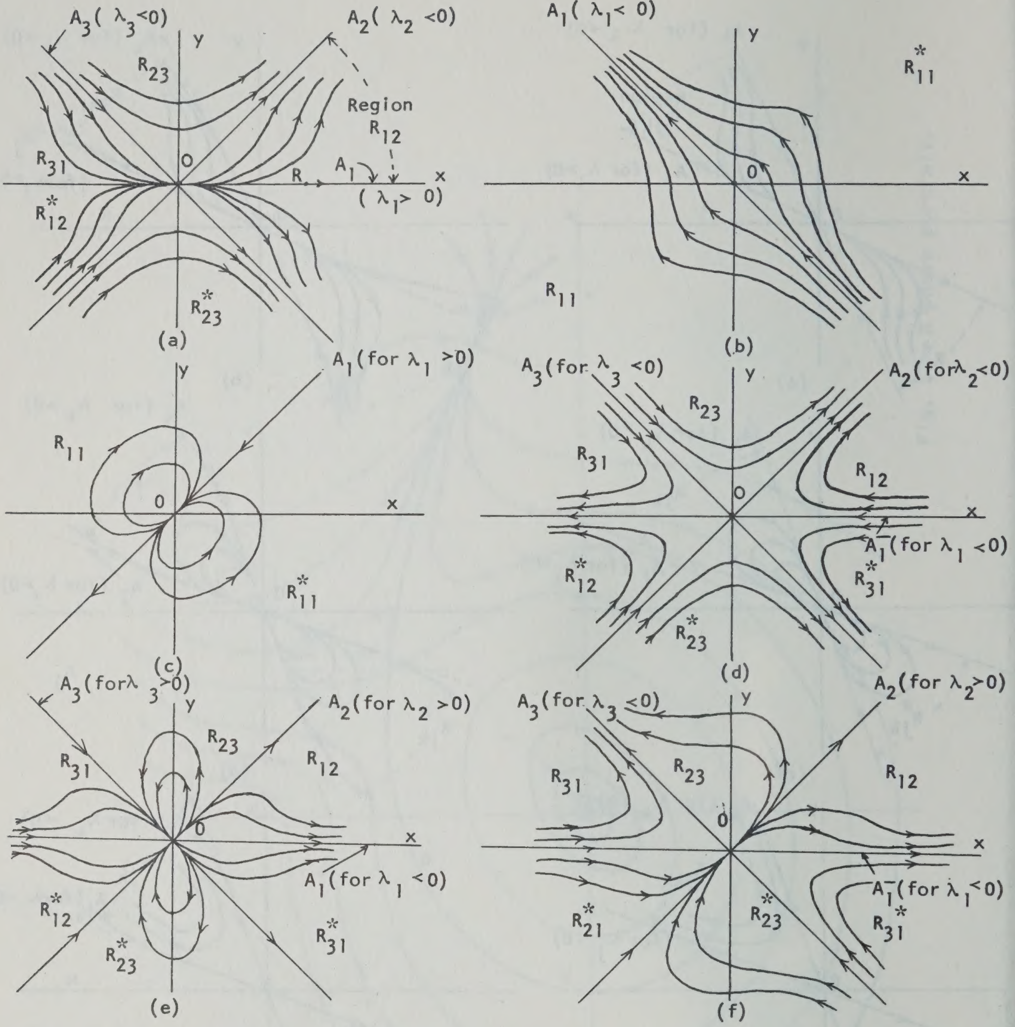

y
A_k (for $\lambda_k < 0$)
Region
R_{jk}
A_j (for $\lambda_j > 0$)
0
x
R^*_{jk}
(a)
y
A_k (for $\lambda_k < 0$)
R_{jk}
A_j
(for $\lambda_j > 0$)
0
x
R^*_{jk}
(b)
y
A_k (for $\lambda_k > 0$)
R_{jk}
A_j (for $\lambda_j > 0$)
0
x
R^*_{jk}
(c)
y
A_k (for $\lambda_k > 0$)
R_{jk}
A_j (for $\lambda_j > 0$)
0
x
A'_j
R^*_{jk}
(d)
y
A_k (for $\lambda_k < 0$)
R_{jk}
A_j
(for $\lambda_j < 0$)
0
x
R^*_{jk}
(e)
y
A_k (for $\lambda_k < 0$)
R_{jk}
A_j (for $\lambda_j < 0$)
x
R^*_{jk}
(f)

Fig. 3 - (a), (b) Open nodal regions. (c), (d) Closed nodal regions.
(e), (f) Saddle regions.



$A_3(\lambda_3<0)$
$A_2(\lambda_2<0)$
y
R_{23}
Region
R_{12}
R_{31}
O
R
A_1
x
$(\lambda_1>0)$
R_{12}^*
R_{23}^*
(a)
$A_1(\lambda_1<0)$
y
R_{11}^*
0
x
R_{11}
(b)
A_1(for $\lambda_1>0$)
y
R_{11}
0
x
R_{11}^*
(c)
A_3(for $\lambda_3<0$)
y
A_2(for $\lambda_2<0$)
R_{23}
R_{31}
R_{12}
O
x
A_1^-(for $\lambda_1<0$)
R_{12}^*
R_{31}^*
R_{23}^*
(d)
A_3(for $\lambda_3>0$)
y
A_2(for $\lambda_2>0$)
R_{23}
R_{31}
R_{12}
0
x
A_1(for $\lambda_1<0$)
R_{12}^*
R_{23}^*
R_{31}^*
(e)
A_3(for $\lambda_3<0$)
y
A_2(for $\lambda_2>0$)
R_{31}
R_{23}
R_{12}
0
x
A_1^-(for $\lambda_1<0$)
R_{21}^*
R_{23}^*
R_{31}^*
(f)

Fig. 4 - Some non-elementary singular points for the system in (8) with n = 2 having characteristic numbers, $[\alpha_1, \alpha_2, \alpha_3; \lambda_1, \lambda_2, \lambda_3; \bar{P}_2(1, \alpha_1)]$ or $[\alpha_1; \lambda_1; \bar{P}_2(1, \alpha_1)]$ with complex α_2, α_3.

(a) [0,1,-1; (+),(-),(-);(+)] (b) [-1;(-);(-)] . (c) [1;(+);(-)] .
(d) [0,1,-1;(-),(-),(-);(-)]. (e) [0,1,-1;(-),(+),(+);(+)].
(f) [0,1,-1;(-),(+),(-);(+)]

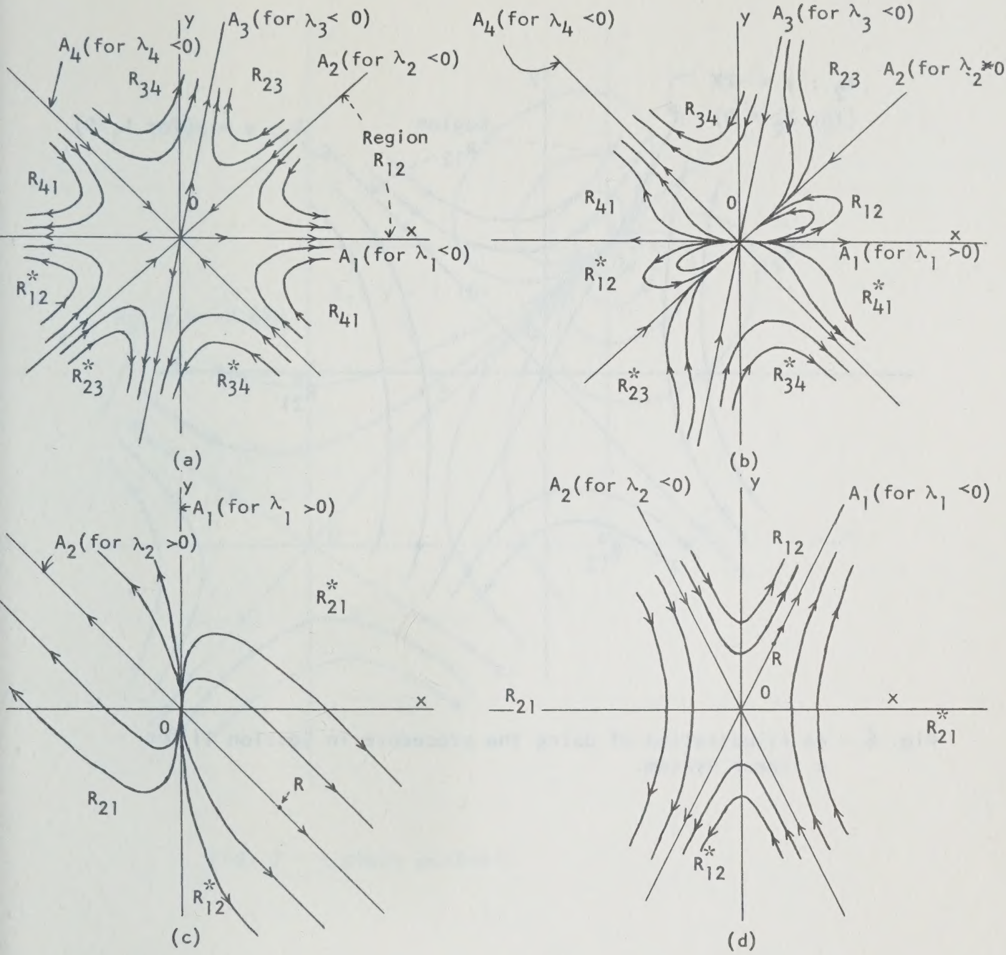

y
A_3(for $\lambda_3 < 0$)
A_4(for $\lambda_4 < 0$)
A_2(for $\lambda_2 < 0$)
R_{34}
R_{23}
Region R_{12}
R_{41}
0
x
A_1(for $\lambda_1 < 0$)
R^*_{12}
R_{41}
R^*_{23}
R^*_{34}
(a)
A_4(for $\lambda_4 < 0$)
y
A_3(for $\lambda_3 < 0$)
R_{23}
A_2(for $\lambda_2 > 0$)
R_{34}
R_{41}
0
R_{12}
x
A_1(for $\lambda_1 > 0$)
R^*_{12}
R^*_{41}
R^*_{23}
R^*_{34}
(b)
y
A_1(for $\lambda_1 > 0$)
A_2(for $\lambda_2 > 0$)
R^*_{21}
x
0
R_{21}
R
R^*_{12}
(c)
A_2(for $\lambda_2 < 0$)
y
A_1(for $\lambda_1 < 0$)
R_{12}
R
0
R_{21}
x
R^*_{21}
R^*_{12}
(d)

Fig. 5 - Some non-elementary singular points for the system in (8) with n = 3 having characteristic numbers, $[\alpha_1, \alpha_2, \alpha_3, \alpha_4; \lambda_1, \lambda_2, \lambda_3, \lambda_4; P_2(1, \alpha_1)]$ or $[\alpha_1, \alpha_2; \lambda_1, \lambda_2; P_2(1, \alpha_i)]$ with complex α_3, α_4.

(a) [0,2/3,5,-1;(-),(-),(-),(-);(+)]. (b) [0,2/3,5,-1;(+),(+),(-),(-);(+)]. (c) [∞,-1;(+),(-);(+)] for i = 2. (d) [2,-2;(-),(-);(+)] for i = 1.

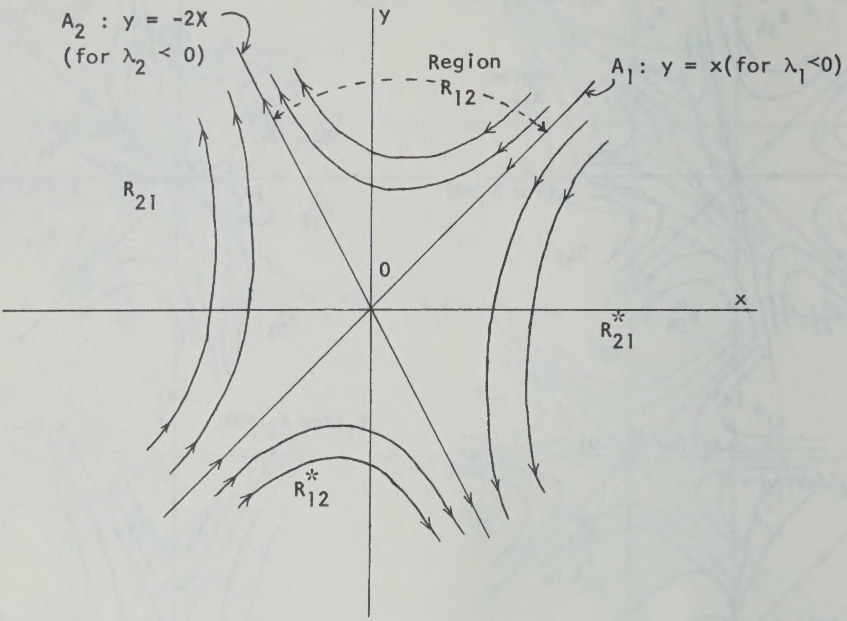

$A_2 : y = -2X$
(for $\lambda_2 < 0$)
y
Region
R_{12}
$A_1 : y = x$(for $\lambda_1 < 0$)
R_{21}
0
x
R^*_{21}
R^*_{12}

Fig. 6 - An illustration of using the procedure in Section II for a linear system.

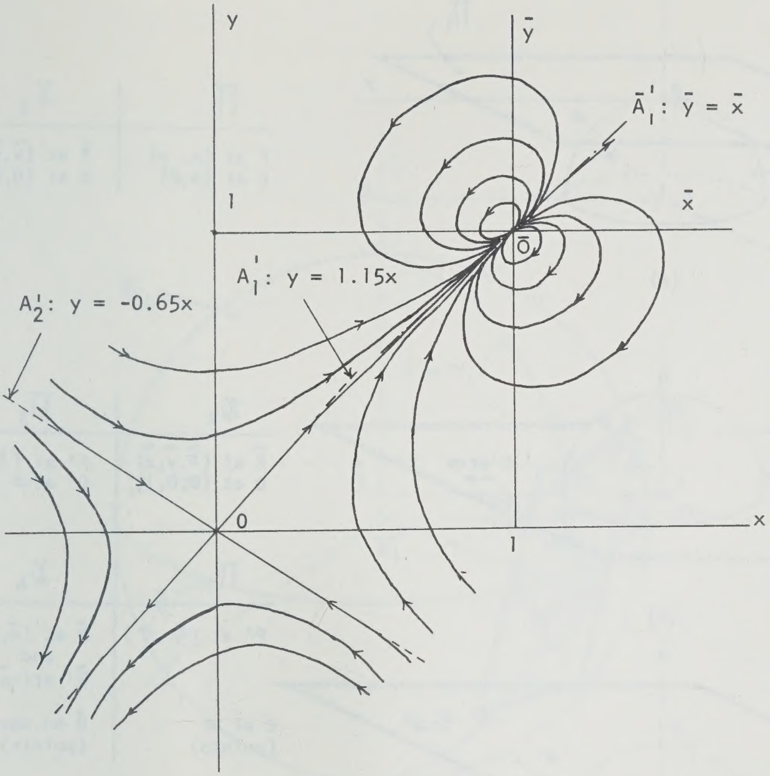
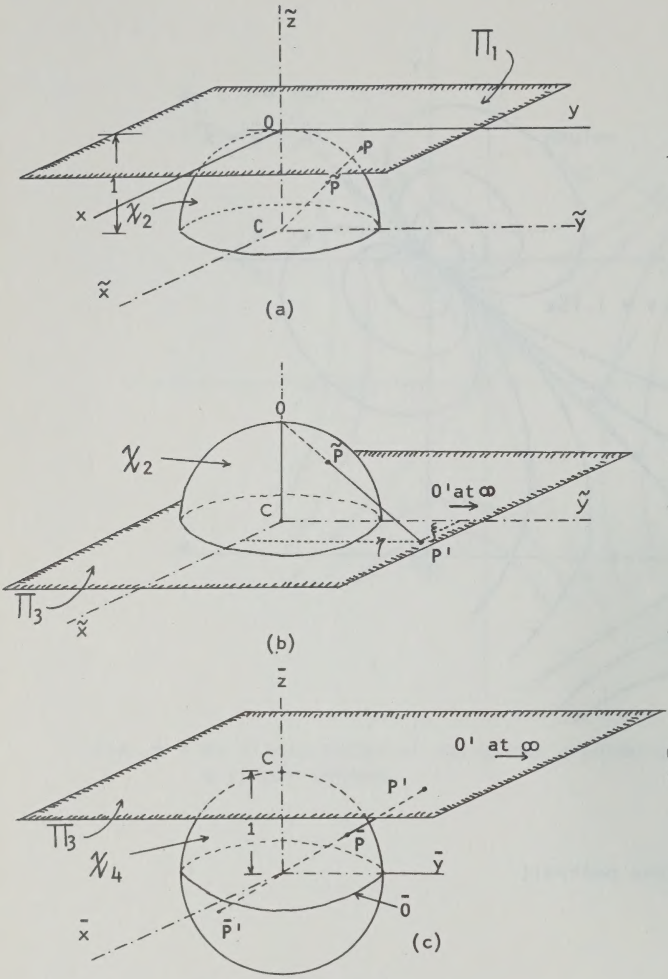

y
ȳ
Ā'₁: ȳ = x̄
1
x̄
Ō
A'₁: y = 1.15x
A'₂: y = -0.65x
0
1
x

Fig. 7 - A phase portrait


$\tilde{z}$
Π_1
0
y
P
$\tilde{P}$
1
x
χ_2
C
$\tilde{y}$
$\tilde{x}$
(a)
0
χ_2
$\tilde{P}$
0' at ∞
C
$\tilde{y}$
η
ξ
P'
Π_3
$\tilde{x}$
(b)
$\bar{z}$
C
0' at ∞
P'
1
Π_3
χ_4
$\bar{P}$
$\bar{y}$
$\bar{0}$
$\bar{x}$
$\bar{P}'$
(c)

Π_1	χ_2
P at (x, y)	$\tilde{P}$ at ($\tilde{x},\tilde{y},\tilde{z}$)
0 at (0,0)	0 at (0,0,1)

χ_2	Π_3
$\tilde{P}$ at ($\tilde{x},\tilde{y},\tilde{z}$)	P' at (ξ , η)
0 at (0,0,1)	0' at ∞ (points)

Π_3	χ_4
P' at (ξ , η)	$\bar{P}$ at ($\bar{x},\bar{y},\bar{z},$) and $\bar{P}'$ at($-\bar{x},-\bar{y},-\bar{z}$)
0 at ∞ (points)	$\bar{0}$ on equator (points)

Fig. 8 - Transformation $M = T_3T_2T_1$ that maps the origin 0 of the Π_1-plane into the equator $\bar{0}$ of the χ_4- plane.

(a) T_1 that maps Π_1 unto χ_2. (b) T_2 that maps χ_2 unto Π_3.

(c) T_3 that maps Π_3 unto χ_4.

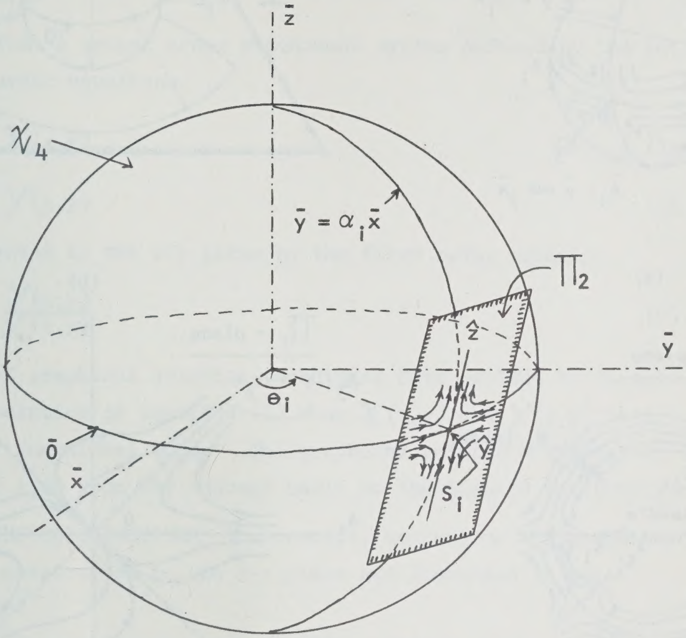

$\bar{z}$
χ_4
$\bar{y} = \alpha_i \bar{x}$
Π_2
$\hat{z}$
$\bar{y}$
θ_i
$\bar{O}$
$\bar{x}$
$\hat{y}$
S_i

Fig. 9 - Investigation of elementary singular points on the equator $\bar{O}$ of χ_4.

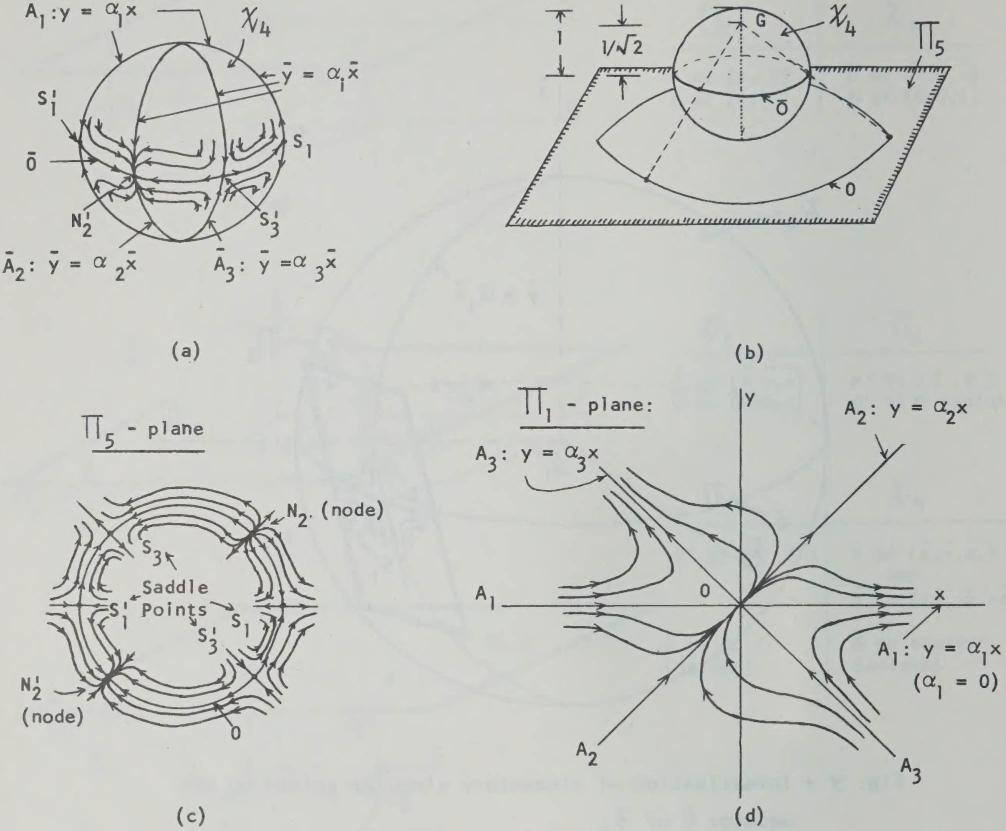

Ā₁ : ȳ = α₁x̄
χ₄
ȳ = α₁x̄
S'₁
S₁
Ō
N'₂
S'₃
Ā₂ : ȳ = α₂x̄
Ā₃ : ȳ = α₃x̄
(a)
1
1/√2
G
χ₄
Π₅
Ō
0
(b)
Π₅ - plane
N₂ (node)
S₃
Saddle Points
S'₁
S₁
S'₃
N'₂ (node)
0
(c)
Π₁ - plane:
y
A₂: y = α₂x
A₃: y = α₃x
A₁
0
x
A₁: y = α₁x
(α₁ = 0)
A₂
A₃
(d)

Fig. 10 - (a) Phase representation in the neighborhood of the equator $\bar{0}$ of the sphere χ_4. (b) Transformation T_4 that maps χ_4 unto Π_5.

(c) The phase representation in the neighborhood of a circle 0 in Π_5.

(d) The phase portrait in the neighborhood of the origin 0 in the phase plane Π_1.

SECOND ORDER AUTONOMOUS, NONLINEAR SYSTEM
A GRAPHICAL SOLUTION

Theodore A. Bickart
Electrical Engineering Department
Syracuse University, Syracuse, New York

INTRODUCTION A second order autonomous system defined by the pair of first order equations

$$\frac{dx}{dt} = \mathcal{X}(x,y) \qquad (1a)$$

$$\frac{dy}{dt} = \mathcal{Y}(x,y) \qquad (1b).$$

is represented in the x-y plane by the first order equation

$$\frac{dy}{dx} = \frac{\mathcal{Y}(x,y)}{\mathcal{X}(x,y)} . \qquad (2)$$

A number of graphical solution techniques have evolved to facilitate the solution of Equation (2) when $\mathcal{X}(x,y)$ and $\mathcal{Y}(x,y)$ have particular functional forms. The graphical solutions treated in this paper stem from the concept basic to the Lienard construction.

LIENARD-LIKE CONSTRUCTIONS The concept, central to the graphical solutions constructed in the x-y plane and discussed in this section, is:

> Given P, a solution point of (2) in the x-y plane, the construction is required to locate a point A such that the line $\overline{AP}$ is normal to the solution curve at P. The solution curve is then extended to the point P', which is a short distance from P in a direction normal to $\overline{AP}$. (P' replaces P in the repetition of this process.)

The concept, as illustrated in Figure 1, is simple; it is the construction to locate point A which proved to be so elusive in the past.

The following survey of Lienard-like constructions indicates the particular form of (2) for which graphical solution techniques have been devised. The coordinates of A are immediately evident, for, if point P is (x_o,y_o), it is a simple matter to show that

A must have $x_o+\mathcal{Y}(x_o,y_o)$ as its x-coordinate and $y_o-\mathcal{X}(x_o,y_o)$ as its y-coordinate (or $x_o-\mathcal{Y}(x_o,y_o)$ as its x-coordinate and $y_o+\mathcal{X}(x_o,y_o)$ as its y-coordinate).

Lienard[1,2] presented the construction to locate point A when (2) is of the form

$$\frac{dy}{dx} = - \frac{x}{y+Y(x)}. \tag{3}$$

The obviously similar construction for

$$\frac{dy}{dx} = - \frac{x+X(y)}{y} \tag{4}$$

is often presented as the Lienard construction in textbooks.[3,4,5,6,7] Clauser[8] extended the Lienard construction when he observed that the constructions for (3) and (4) could be combined to provide the needed construction for

$$\frac{dy}{dx} = - \frac{x+X(y)}{y+Y(x)}. \tag{5}$$

The construction was further generalized by Bickart[9] to permit solution of

$$\frac{dy}{dx} = - \frac{x + X}{y + Y} \tag{6}$$

where X may be a function of x or y and Y may be a function of x or y.

Pell[10,11] made a significant departure in the technique employed to locate point A in the Lienard-like construction for

$$\frac{dy}{dx} = - \frac{Y_1(y) + Y_2(x)}{y}. \tag{7}$$

Obviously the construction for

$$\frac{dy}{dx} = - \frac{x}{X_1(y) + X_2(x)} \tag{8}$$

is similar. After combining the constructions for (7) and (8), it is possible to graphically solve

$$\frac{dy}{dx} = - \frac{Y_1(y) + Y_2(x)}{X_1(y) + X_2(x)}. \tag{9}$$

The construction for (9) has been discussed by Reis[12] when $X_2(x) = 0$ and by Bickart[13] when no conditions are imposed.

The summary of constructions presented thusfar provides the procedures for graphically solving (2) when $\mathcal{X}(x,y)$ and/or $\mathcal{Y}(x,y)$ can be expressed as a sum of functions, with each function a function of x or y. It has been shown by Bickart[14] that products can also be obtained graphically, thereby permitting solution of

$$\frac{dy}{dx} = - \frac{x + Y_1(y)\,Y_2(x)}{y} \tag{10}$$

or

$$\frac{dy}{dx} = - \frac{x}{y + X_1(y)\,X_2(x)} \tag{11}$$

and, thence

$$\frac{dy}{dx} = - \frac{x + Y_1(y)\,Y_2(x)}{y + X_1(y)\,X_2(x)}. \tag{12}$$

A construction for

$$\frac{dy}{dx} = - \frac{x + F_1[Y_1(y)\,Y_2(x)]}{y + F_2[X_1(y)\,X_2(x)]} \tag{13}$$

is an obvious consequence of the previous constructions.

The conclusion to be reached after surveying the field of Lienard-like constructions for (2) is: A construction can be specified which locates point A whenever $\mathcal{X}(x,y)$ and $\mathcal{Y}(x,y)$ can be expressed as a combination of simple operations -- obtaining a sum, a product, and a function. It is not necessary to separately specify the product operation, since it will be shown to be a combination of the sum and function operations. The ensuing comments in this section provide a survey of procedures for performing these operations graphically.

First, let us examine the problem of obtaining a point having as its coordinates, functions of either variable of another point. Starting from the point (x_o,y_o), it is relatively easy to locate the point $(G_1(y_o), G_2(x_o))$; this operation is illustrated in Figure 2. Since this operation may be employed to locate the point (y_o,x_o), it is obvious that a second step may be employed to locate the point $(H_1(x_o),H_2(y_o))$. Now that it is possible to locate a point with an x-coordinate which is a desired function of x_o or y_o and another point with a y-coordinate which is a desired function of x_o or y_o, it is a trivial step to locate a resultant point having both coordinates as desired functions of either of the coordinates of the starting point.

Next, let us examine some of the procedures available for obtaining sums (and differences). Commencing with the point (x_1,y_1), the points $(x_1 \pm y_1,0)$ and $(0,y_1 \pm x_1)$ are located by the operation shown in Figure 3. In conjunction with the operation of obtaining a function, the points $(y_1 \pm x_1,0)$ and $(0,x_1 \pm y_1)$ are readily located; in fact, this may be done in a number of ways.

If the coordinates of the point (x_1,y_1) are functions of the coordinates of the point (x_o,y_o), then the demonstrated techniques make it possible to locate a point having an x- or y-coordinate which is the sum (or difference) of two functions of the coordinates of the point (x_o,y_o). Quite obviously, a point can now be found for which both the x- and y-coordinates are sums (or differences) of two functions of the coordinates of another point.

It is now apparent that if the coordinates of the point A, in the Lienard-like construction, can be expressed as a combination of the two, discussed operations on the coordinates of the point P, then a construction for the graphical solution of (2) can be devised.

In designing a "best" construction for any one particular problem, alternate summing procedures will be valuable. A parallel line construction for adding (subtracting) the x-coordinate of one point, (x_1,y_1), to (from) the x-coordinate of a second point, (x_2,y_2), is illustrated in Figure 4. A similar construction may be applied to the y-coordinates. Thus, separately or in combination these constructions make the points $(x_2 \pm x_1, y_2 \pm y_1)$, $(x_2 \pm x_1,y_2)$, and $(x_2, y_2 \pm y_1)$ accessible. If the summing construction illustrated in Figure 3 is carried out relative to a set of axes centered at the point (x_2,y_2), then the points located are $(x_1 \pm (y_1-y_2),y_2)$ and $(x_2,y_1 \pm (x_1-x_2))$; this has been shown in Figure 5. This illustrates one of a number of similar constructions for performing a triple sum (with selected terms reversed in sign).

It is indeed interesting to realize that the operations considered thusfar are sufficient to effect a construction which will multiply the coordinates of the point (x_1,y_1) and then locate the point $(x_1y_1,0)$ or $(0,x_1,y_1)$. This is possible through implementation of the quarter-square multiplication technique, which may be expressed as:

$$\frac{1}{4}\left\{[x_1 + y_1]^2 - [x_1 - y_1]^2\right\} = x_1y_1. \tag{14}$$

The construction shown in Figure 6 uses one parabola to locate the points $(x_1+y_1, \frac{1}{4}[x_1+y_1]^2)$ and $(x_1-y_1, \frac{1}{4}[x_1-y_1]^2)$. After finding the point $(0, \frac{1}{4}[x_1+y_1]^2)$, a parallel line construction (see Figure 4) locates $(0,x_1y_1)$. A different technique could have been employed to take the difference of the squares, called for by (14).

In order to provide a distinctly different construction for multiplication, it is first necessary to consider another construction "trick." As shown in Figure 7, the x'-y' plane is obtained by rotating the x-y plane through an angle Θ. Thus:

$$x' = \cos\Theta\, x + \sin\Theta\, y \tag{15a}$$
$$y' = -\sin\Theta\, x + \cos\Theta\, y \tag{15b}$$

and

$$x = \cos\Theta\, x' - \sin\Theta\, y' \tag{16a}$$
$$y = \sin\Theta\, x' + \cos\Theta\, y' \tag{16b}$$

A construction carried out in the x'-y' plane according to the techniques discussed previously may be considerably simpler than the construction in the x-y plane needed to locate the same point.

A second multiplication scheme makes use of the fact that for $\Theta = -45^{\circ}$, $x' = [x-y]/\sqrt{2}$ and $y' = [x+y]/\sqrt{2}$; that is, x' and y' are proportional to the sum and difference of x and y. The point (x_1,y_1) becomes $(x_1',y_1') = ([x_1-y_1]/\sqrt{2}, [x_1+y_1]/\sqrt{2})$ in the x'-y' plane. Now in the x'-y' plane the point $(-y_1'^2/\sqrt{2}, -x_1'^2/\sqrt{2}) = (-[x_1+y_1]^2/2\sqrt{2}, -[x_1-y_1]^2/2\sqrt{2})$ is very easy to locate. Then, (16b) indicates that, in the x-y plane, the y-coordinate of that point is x_1y_1. This two parabola construction to locate the x-y plane point $(0,x_1y_1)$ is illustrated in Figure 8.

Both multiplication schemes discussed have determined a point with its y-coordinate equal to the product of the coordinates of another point. Similar constructions will locate a point with its x-coordinate being that product.

Many of the constructions presented for the operations of determining a function, finding a sum, and obtaining a product have been employed as parts of the complete constructions for graphically solving (3) through (13), as presented in the cited literature. Numerous, graphically solved problems have been presented in the referenced works; the reader is referred to these sources to see the application of the constructions discussed in this section.

A fitting observation with which to end this section is: The constructions presented for the three "basic" operations are independent of the concept central to the Lienard-like constructions and are, therefore, quite general. The implication is that these basic constructions might be used in graphically solving problems according to a different concept.

GENERALIZED CONSTRUCTION CONCEPT The concept which is central to the Lienard-like constructions is restrictive in the following sense: The solution curve is extended from P to P' in a direction normal to $\overline{AP}$ -- in a direction 90° from the direction of $\overline{AP}$ and not any other angle. Let us, therefore, generalize the construction concept. This generalized concept, illustrated in Figure 9, may be stated as follows:

> Given P, a solution point of (2) in the x-y plane, the construction is required to locate a point A_α such that the angle between $\overline{A_\alpha P}$ and the solution curve at P is α. The solution curve is then extended to the point P' which

is a short distance from P in a direction which forms an angle α with $\overline{A_\alpha P}$. (P' replaces P in the repetition of this process.)

Now let us determine what the functions $\mathcal{X}(x,y)$ and $\mathcal{Y}(x,y)$ in (2) should be like to effectively employ this construction concept. Let the point P be (x_o,y_o) and the point A_α be $(F_1(x_o,y_o), F_2(x_o,y_o))$. Then

$$\text{slope } \overline{A_\alpha P} = \frac{y_o - F_2(x_o,y_o)}{x_o - F_1(x_o,y_o)} \tag{17}$$

and

$$\text{slope } \overline{PP'} = \frac{\text{slope } \overline{A_\alpha P} - \tan\alpha}{1+\text{slope } \overline{A_\alpha P}\ \tan\alpha}$$

$$= \frac{\cos\alpha\,[y_o-F_2(x_o,y_o)] - \sin\alpha\,[x_o-F_1(x_o,y_o)]}{\cos\alpha\,[x_o-F_1(x_o,y_o)] + \sin\alpha\,[y_o-F_2(x_o,y_o)]} \tag{18}$$

but, at (x_o,y_o), $dy/dx = \text{slope } \overline{PP'}$. Thus, since these relations are to hold for any (x_o,y_o), Equation (2) will be of the form

$$\frac{dy}{dx} = \frac{a\ G_2(x,y) - b\ G_1(x,y)}{a\ G_1(x,y) + b\ G_2(x,y)} \tag{19}$$

with $\cos\alpha = a/(a^2+b^2)^{1/2}$, $\sin\alpha = b/(a^2+b^2)^{1/2}$, $F_1(x,y) = x-G_1(x,y)$ and $F_2(x,y) = y - G_2(x,y)$. We consequently find that $\mathcal{Y}(x,y) = a\ G_2(x,y) - b\ G_1(x,y)$ and $\mathcal{X}(x,y) = a\ G_1(x,y) + b\ G_2(x,y)$.

When $a = 0$ $(\alpha = 90^\circ)$, the Lienard-like constructions are obtained. When $b = 0$ $(\alpha = 0^\circ)$, (19) becomes

$$\frac{dy}{dx} = \frac{G_2(x,y)}{G_1(x,y)} \tag{20}$$

which is similar in form, with the exception of a minus sign, to (19) with $a = 0$:

$$\frac{dy}{dx} = -\frac{G_1(x,y)}{G_2(x,y)}. \tag{21}$$

For (20) the solution curve is projected to P' along $\overline{A_{0^o}P}$, this eliminates the construction of a right angle -- the angle needed when projecting the solution curve to P' for (21). This assumes that A_{0^o} and A_{90^o} are located by constructions of comparable difficulty. When (2) is of the form in (9), which is that of (21) with $G_1(x,y) = Y_1(y) + Y_2(x)$ and $G_2(x,y) = X_1(y) + X_2(x)$ or that of (20) with $G_1(x,y) = X_1(y) + X_2(x)$ and $G_2(x,y) = -Y_1(y) - Y_2(x)$, the constructions to locate A_{0^o} and A_{90^o} would be similar; hence, it would be advantageous to employ the construction for $\alpha = 0^o$.

To demonstrate that the generalized construction concept has value for $\alpha \neq 0^o$ or 90^o, consider that the system described by

$$\frac{dx}{dt} = -\gamma x + y \qquad \text{and} \qquad \frac{dy}{dt} = -x - \gamma y \tag{22}$$

has the x-y plane equation

$$\frac{dy}{dx} = \frac{-x - \gamma y}{y - \gamma x} \tag{23}$$

which is (19) with $G_2(x,y) = -x$, $G_1(x,y) = y$, $a = 1$, and $b = \gamma$. Note, of course, that (22) are the parametric equations for a logarithmic spiral in the x-y plane.

CONCLUDING DISCUSSION An extensive development of graphical solution techniques stemming originally from Lienard's construction has provided the means for obtaining a "phase" plane solution of the describing equations of an exceedingly large class of second order, autonomous systems. It is important to recognize, though, that other procedures are available for obtaining graphical solutions of (2).

Jacobsen's δ-method[15,16,17,18,19,20] employs the same concept as the Lienard-like constructions, with the exception that the solution curve is extended from P to P' as the arc of a circle centered at A. It is possible to modify the δ-method[21] so as to get slope and curvature fitting of the graphical solution approximation to the solution; this is an advantage over the Lienard-like graphical solutions which only fit the slope.

The acceleration-plane method due to Ku[22,23,24] follows the same concept as in the δ-method to obtain a solution of (2) when of the form

$$\frac{dy}{dx} = \frac{\mathcal{Y}(x,y)}{y}. \tag{24}$$

The procedure for locating the point A is not exact; graphical interpolation is necessary.

The slope-line method[25,26,27] for solution of (2) in the form of (5) has the distinct advantage of providing the time corresponding to every constructed point on the solution curve. The construction to obtain a solution by the slope-line method is, however, quite different from the constructions discussed in this paper.

The method of isoclines[28,29,30,31] is a very tedious technique for obtaining a solution to (2). A somewhat similar method, due to LeCorbeiller,[32] greatly reduces the time needed to obtain a solution when (2) is of the form

$$\frac{dy}{dx} = -\frac{x + X(x,y)}{y} \tag{25}$$

or the form

$$\frac{dy}{dx} = -\frac{x}{y + Y(x,y)}. \tag{26}$$

The list of references is a comprehensive bibliography of basic and original sources for the graphical solution of (2).

REFERENCES

1. A. Lienard, "Etude des oscillations entretenues," Revue Generale de L'Electricite, Vol. 23, pp. 901-912, 946-954, May 1928.
2. Nicholas Minorsky, Nonlinear Oscillations, D. Van Nostrand Company (Princeton, N.J., 1962), pp. 108-110.
3. J.J. Stoker, Nonlinear Vibrations, Interscience Publishers (New York, 1950), pp. 31-32.
4. George J. Thaler and Marvin P. Pastel, Analysis and Design of Nonlinear Feedback Control Systems, McGraw-Hill Book Company (New York, 1962), pp. 71-72.
5. Y.H. Ku, Analysis and Control of Nonlinear Systems, Ronald Press (New York, 1958), pp. 128-130.

6. Dunstan Graham and Duane McRuer, Analysis of Nonlinear Control Systems, John Wiley & Sons (New York, 1961), pp. 302-304.

7. William Lewis Hughes, Nonlinear Electrical Networks, Ronald Press (New York, 1960), pp. 178-180.

8. Francis H. Clauser, "The Behavior of Nonlinear Systems," Journal of Aeronautical Sciences, Vol. 23, No. 5, pp. 411-434, May 1956

9. Theodore A. Bickart, "Extension of the Lienard Construction," IEEE Trans. on Circuit Theory, Vol. CT-11, No. 2, pp. 288-289, June 1964.

10. W.H. Pell, "Graphical Solution of Single-Degree-of-Freedom Vibration Problem with Arbitrary Damping and Restoring Forces," Journal of Applied Mechanics, pp. 311-312, June 1957.

11. John E. Gibson, Nonlinear Automatic Control, McGraw-Hill Book Company (New York, 1963) pp. 244-246.

12. G. Reis, "An Extension of Pell's Method of Phase-Trajectory Construction," IEEE Trans. on Automatic Control, Vol. AC-9, No. 3, p. 315, July 1964.

13. Theodore A. Bickart, "Comments Following From 'An Extension of Pell's Method of Phase-Trajectory Construction,'" Syracuse Univ. Dept. of Elec. Engg. Technical Memorandum TM-64-8, August 1964 and IEEE Trans. on Automatic Control, Vol. AC-10, April 1965.

14. Theodore A. Bickart, "Nonlinear Analysis in the Phase Plane," Journal of the Franklin Institute, October 1964.

15. L.S. Jacobsen, "On a General Method of Solving Second-Order Ordinary Differential Equations by Phase-Plane Displacements," Journal of Applied Mechanics, pp. 543-553, December 1952.

16. Robert N. Buland, "Analysis of Nonlinear Servos by Phase-Plane-Delta Method," J. of Franklin Inst., Vol. 257, pp. 37-48, Jan'54

17. W.J. Cunningham, Introduction to Nonlinear Analysis, McGraw-Hill Book Company (New York, 1958) pp. 39-44.

18. John E. Gibson, op. cit., pp. 246-249.

19. George J. Thaler and Marvin P. Pastel, op. cit., pp. 72-76.

20. Y.H. Ku, Analysis and Control of Nonlinear Systems, op. cit. pp. 37-38.

21. Theodore A. Bickart, "Graphical Analysis of Nonlinear Systems: A Modification of the δ-Method," IEEE Trans. on Education, (to be published).

22. Y.H. Ku, "Nonlinear Analysis of Electro-Mechanical Problems," J. of Franklin Inst., Vol. 255, pp. 9-31, January 1953.

23. Y.H. Ku, "A Method for Solving Third and Higher Order Nonlinear Differential Equations," J. of Franklin Inst., Vol. 256, pp. 229-244, September 1953.

24. Y.H. Ku, Analysis and Control of Nonlinear Systems, op. cit., pp. 21-24.

25. Henry M. Paynter, "Methods and Results from MIT Studies in Unsteady Flow," Journal of the Boston Society of Civil Engineers, Vol. 39, pp. 120-165, 1952.

26. Y.H. Ku, "Analysis and Control of Nonlinear Systems," op. cit, pp. 38-39.

27. R.H. MacMillan, Nonlinear Control Systems Analysis, Pergamon Press (Oxford, 1962), pp. 98-116.

28. W.J. Cunningham, op. cit., pp. 28-36.

29. Dunstan Graham and Duane McRuer, op. cit, pp. 295-300.

30. John E. Gibson, op. cit., pp. 238-244.

31. George J. Thaler and Marvin P. Pastel, op. cit., pp. 68-70.

32. P. LeCorbeiller, "Two-Stroke Oscillators," IRE Trans. on Circuit Theory, Vol. CT-7, No. 4, pp. 387-398, December 1960.

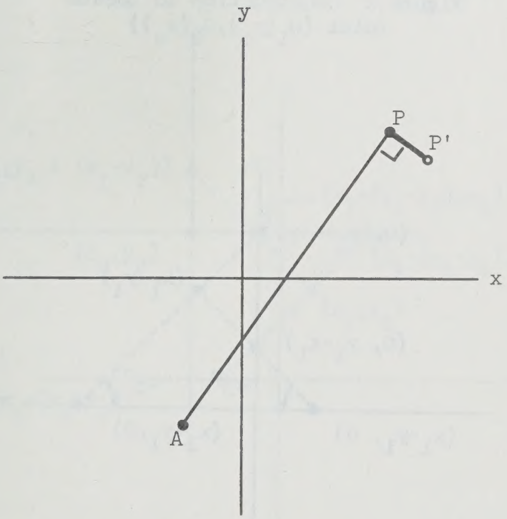


Figure 1 Illustration of Concept Basic to Lienard-Like Constructions

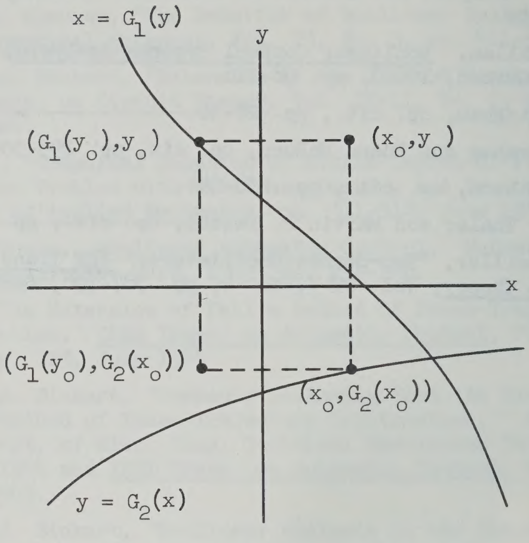

$x = G_1(y)$
y
$(G_1(y_o), y_o)$
(x_o, y_o)
x
$(G_1(y_o), G_2(x_o))$
$(x_o, G_2(x_o))$
$y = G_2(x)$

Figure 2 Construction to Locate Point $(G_1(y_o), G_2(x_o))$

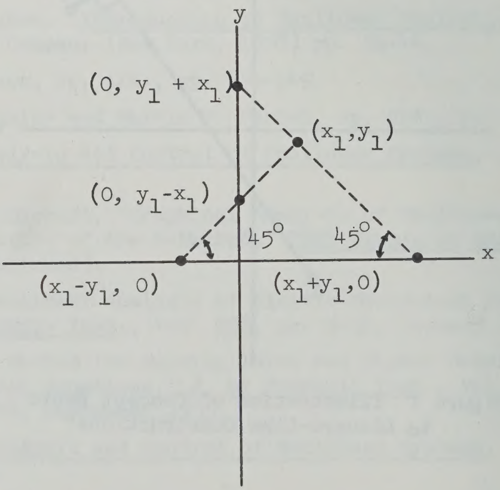

y
$(0, y_1 + x_1)$
(x_1, y_1)
$(0, y_1 - x_1)$
45^o
45^o
x
$(x_1 - y_1, 0)$
$(x_1 + y_1, 0)$

Figure 3 Construction to Locate Points $(x_1 \pm y_1, 0)$ and $(0, y_1 \pm x_1)$

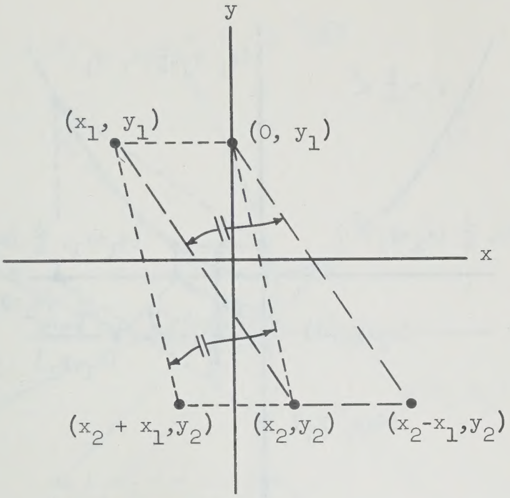

y
(x_1, y_1)
$(0, y_1)$
x
$(x_2 + x_1, y_2)$
(x_2, y_2)
$(x_2 - x_1, y_2)$

Figure 4 Construction to Locate Points $(x_2 \pm x_1, y_2)$

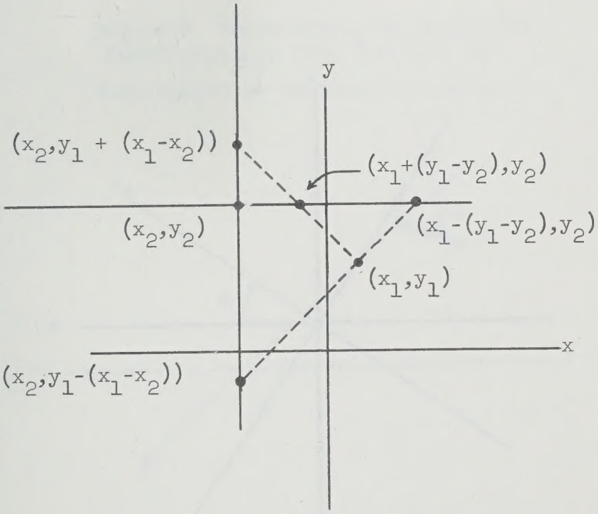

y
$(x_2, y_1 + (x_1 - x_2))$
$(x_1 + (y_1 - y_2), y_2)$
(x_2, y_2)
$(x_1 - (y_1 - y_2), y_2)$
(x_1, y_1)
x
$(x_2, y_1 - (x_1 - x_2))$

Figure 5 Construction to Locate Points $(x_1 \pm (y_1 - y_2), y_2)$ and $(x_2, y_1 \pm (x_1 - x_2))$

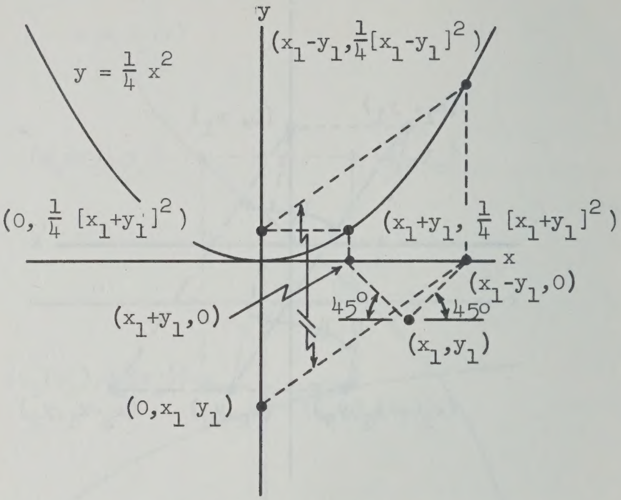

y
$y = \frac{1}{4} x^2$
$(x_1-y_1, \frac{1}{4}[x_1-y_1]^2)$
$(0, \frac{1}{4} [x_1+y_1]^2)$
$(x_1+y_1, \frac{1}{4} [x_1+y_1]^2)$
x
$(x_1-y_1, 0)$
45^o
45^o
$(x_1+y_1, 0)$
(x_1, y_1)
$(0, x_1 y_1)$

Figure 6 Construction to Locate the Point $(0, x_1y_1)$

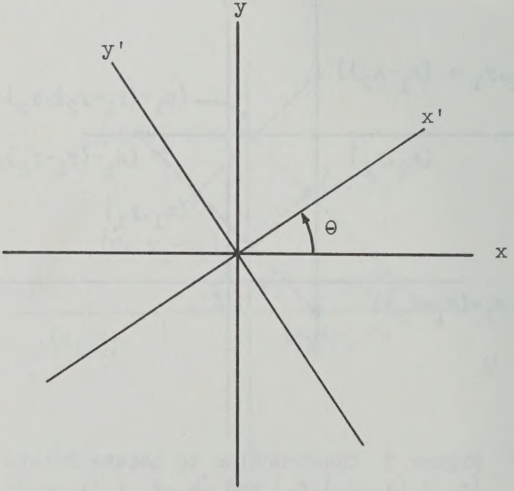


Figure 7 Rotation of x-y Plane Into x'-y' Plane

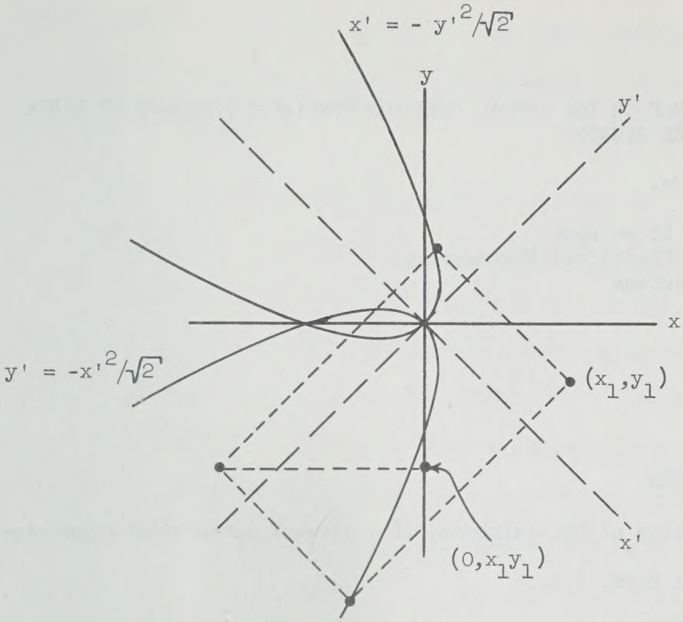

$x' = -y'^2/\sqrt{2}$
y
y'
x
$y' = -x'^2/\sqrt{2}$
(x_1,y_1)
x'
$(0,x_1y_1)$

Figure 8 Construction to Locate the Point $(0,x_1y_1)$ [See the Text For Description of Construction Steps.]

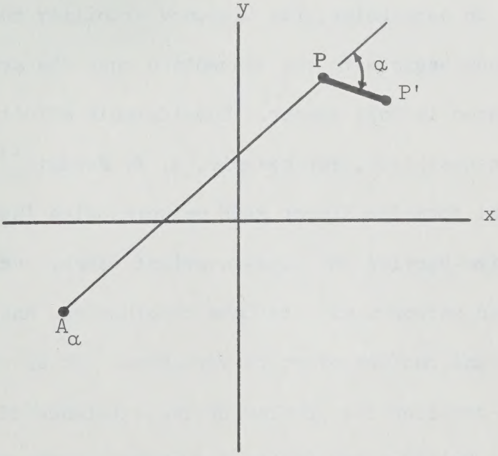

y
P
α
P'
x
A_α

Figure 9 Illustration of Generalized Construction Concept

ON THE EXISTENCE OF THE NORMAL FORM FOR NONLINEAR NETWORKS AT LARGE WITH CONTROLLED SOURCES

Leo V. Auth, Jr.
Ruey-wen Liu
University of Notre Dame
Department of Electrical Engineering
Notre Dame, Indiana

I. INTRODUCTION

The question of the existence of a network mathematical description in normal form, i.e.,

$$x = f(x, t)$$

is of considerable theoretical interest in studying various qualitative properties of network behavior such as stability, boundedness, oscillatory, etc. In particular, the Liapunov stability theorems and their extensions begin with the assumption that the system variables can be formulated in this manner. Considerable effort in this area has recently taken place, for example, P. R. Bryant(1) exhibited an explicit normal form for linear RLCM networks with independent sources in both the time-varying and time-invariant cases. Others(2,3,4,5,6,7,8) have considered networks with various combinations and types of network elements and choices of state variables. It is the purpose of this paper to consider the problem of the existence of the normal form with capacitor voltages and inductor currents as the state variables for a subclass of nonlinear networks at large(6) containing lumped nonlinear time-invariant elements and dependent sources. Such a choice of network variables is pertinent because of their relation to the stored energy and their property of being conveniently observable.

II. NETWORK CONTENT AND CLASSIFICATION OF ELEMENTS

We consider each element of the network N to be a two terminal element. Thus, any sources, whether dependent or not, are considered separately and not lumped together with other series or parallel elements to form "super" branches as in Bryant.(1) Capacitive and inductive elements of N are assumed to be characterized by differentiable functions of voltage and current(s) respectively. To achieve some economy in notation, we do not consider networks with extraneous reactive and source elements, i.e., inductor-only and/or current-source-only cut-sets and capacitor-only and/or voltage-source-only tie-sets are not allowed. In addition, controlled sources are assumed to be dependent on voltages and/or currents and not on their derivatives. Thus, the dimension of the resulting state-vector will be the same as the number of inductive and capacitive elements in the network.

A final restriction on the geometry of the element content of N is that N possesses at least one tree τ such that τ contains all capacitors and voltage sources of N while the cotree of τ contains all inductors and current sources of N. The network elements and relevant coefficient matrices are then classified using subscript notation in the following manner:

α - cotree current sources	ε - tree resistive elements
β – cotree resistive elements	δ – tree capacitive elements

γ - cotree inductive elements ζ - tree voltage sources

III. DIFFERENTIAL-ALGEBRAIC EQUATIONS

Kirchoff's constraint relations for the element voltage-current variables may be expressed as

$$v_\alpha = -F_{\alpha\delta} v_\delta - F_{\alpha\varepsilon} v_\varepsilon - F_{\alpha\zeta} v_\zeta \tag{1}$$

$$v_\beta = -F_{\beta\delta} v_\delta - F_{\beta\varepsilon} v_\varepsilon - F_{\beta\zeta} v_\zeta \tag{2}$$

$$v_\gamma = -F_{\gamma\delta} v_\delta - F_{\gamma\varepsilon} v_\varepsilon - F_{\gamma\zeta} v_\zeta \tag{3}$$

$$i_\delta = F'_{\alpha\delta} i_\alpha + F'_{\beta\delta} i_\beta + F'_{\gamma\delta} i_\gamma \tag{4}$$

$$i_\varepsilon = F'_{\alpha\varepsilon} i_\alpha + F'_{\beta\varepsilon} i_\beta + F'_{\gamma\varepsilon} i_\gamma \tag{5}$$

$$i_\zeta = F'_{\alpha\zeta} i_\alpha + F'_{\beta\zeta} i_\beta + F'_{\gamma\zeta} i_\gamma \tag{6}$$

where the F matrices represent appropriate submatrices of the fundamental circuit matrix with respect to the tree τ. Substitution of the voltage-current relations for the capacitive and inductive elements into Equations (3) and (4) gives

$$L_{\gamma\gamma}(i_\gamma)\dot{i}_\gamma = -F_{\gamma\delta}v_\delta - F_{\gamma\varepsilon}v_\varepsilon - F_{\gamma\zeta}v_\zeta \tag{7}$$

$$C_{\delta\delta}(v_\delta)\dot{v}_\delta = F'_{\gamma\delta}i_\gamma + F'_{\beta\delta}i_\beta + F'_{\alpha\delta}i_\alpha . \tag{8}$$

Therefore, the final formulation of the equation in normal form rests largely upon the question of the existence of the solution to the remaining algebraic equations, i.e., the remaining linear Kirchoff equations and the resistive and source voltage-current relations which may be nonlinear, for the variables i_α , i_β , v_ε , v_ζ in terms of the state variables v_δ and i_γ.

IV. SUFFICIENT CONDITION FOR THE EXISTENCE OF THE NORMAL FORM.

Theorem I: If the elements contained in N satisfy

I) $v_\varepsilon = r(i_\varepsilon)$

II) $i_\beta = g(v_\beta)$

III) $v_\zeta = p_\zeta(v_\delta, i_\gamma) + q_\zeta(v_\beta, i_\varepsilon) + e_\zeta(t)$

IV) $i_\alpha = m_\alpha(v_\delta, i_\gamma) + n_\alpha(v_\beta, i_\varepsilon) + j_\alpha(t)$

then the remaining relevant Kirchoff's Equations (2) and (5) become

$$\begin{aligned} &v_\beta + F_{\beta\varepsilon} r_\varepsilon(i_\varepsilon) + F_{\beta\zeta} q_\zeta(v_\beta, i_\varepsilon) = -F_{\beta\delta} v_\delta - F_{\beta\zeta}[p_\zeta(v_\delta, i_\gamma) \\ &+ e_\zeta(t)] \triangleq z_\beta \end{aligned} \tag{9}$$

$$\begin{aligned} &-F'_{\beta\varepsilon} g(v_\beta) + i_\varepsilon - F'_{\alpha\varepsilon} n_\alpha(v_\beta, i_\varepsilon) = F'_{\gamma\varepsilon} i_\gamma + F'_{\alpha\varepsilon}[m_\alpha(v_\delta, i_\gamma) \\ &+ j_\alpha(t)] \triangleq z_\varepsilon . \end{aligned} \tag{10}$$

These algebraic equations will possess a unique solution for (v_β, i_ε) in terms of (v_δ, i_γ) if

V) The entire (v_β, i_ε) - space maps onto the entire (z_ε, z_β) - space

VI) $||(v_\beta, i_\varepsilon)|| \to \infty \Longleftrightarrow ||(z_\varepsilon, z_\beta)|| \to \infty$

VII) The Jacobian of the left-hand-side of Equations (9) and (10) considered together is not equal to zero.

Finally, Equations (7) and (8) will yield the normal form if:

VIII) $L_{\gamma\gamma}(i_\gamma)$ and $C_{\delta\delta}(v_\delta)$ are definite throughout the state space.

The proof of Theorem I follows immediately from Theorem II which is a consequence of a result given by Palais.(9)

Theorem II: Let X and Y be R^n and $x \varepsilon$ X and $y \varepsilon$ Y.

If: 1) f is a continuous mapping such that

$$f: X \xrightarrow{\text{onto}} Y$$

2) $\lim\limits_{||x|| \to \infty} ||f(x)|| \to \infty$

3) at every point of X, f is a local homeomorphy

then f is a global homeomorphy between X and Y, i.e., f is one-to-one and bi-continuous.

V. ACKNOWLEGEMENT

The authors wish to thank Professor C. A. Desoer for information and suggestions.

VI. REFERENCES

1. P. R. Bryant, "The Explicit Form of Bashkow's A Matrix," IRE Trans. on Circuit Theory, vol. CT-9, pp. 303-306, September, 1962.
2. A. Dervisoglu, "Bashkow's A Matrix for Active RLC Networks," Report R-175, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, July, 1963.
3. R. K. Brayton and J. K. Moser, "A Theory of Nonlinear Networks," Quarterly of Applied Mathematics, pp. 1-33, April, 1964; pp. 81-104, July, 1964.
4. C. A Desoer and J. Katzenelson, "Nonlinear RLC Networks," to appear in Bell System Technical Journal, January, 1965.
5. R. Liu and L. V. Auth, "Qualitative Synthesis of Nonlinear Networks," Proceedings of the First Annual Allerton Conference on Circuit and System Theory, pp. 330-343, November, 1963.
6. R. Liu and L. V. Auth, "Applications of Liapunov's Direct Method to Nonlinear Networks at Large," Technical Report No. EE-642, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, March 3, 1964.
7. L. O. Chua and R. A. Rohrer, "Nonlinear Network Analysis - The Parametric Approach," Report R-224, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, July, 1964.
8. T. E. Stern, "On the Equations of Nonlinear Networks," Paper C9-4 ICMCI, Tokyo, 1964.
9. R. S. Palais, "Natural Operations on Differential Forms," Trans. of the American Mathematical Society, pp. 125-141, July, 1959.

LINEAR DEGENERACIES IN N-PORT NETWORKS

Dr. Ing. Pier Paolo Civalleri
Instituto Elettrotecnico
Nazionale Galileo Ferraris
Torino, Italy

I. INTRODUCTION

This paper deals with the problem of realizing a single-element-kind network exhibiting a performance at n properly chosen terminal pairs (n-port).

The difficulties encountered by various authors who have dealt with such a task arise from the fact that the synthesis of n-ports involves two kinds of problems:

a) the topological problem of finding out a network structure, the properties of which agree with the requested performance;

b) the physical problem of deciding whether the latter can be realized on such a structure using neither non-conventional network elements, such as negative resistors, nor ideal transformers.

The classical methods of network synthesis, dealing almost exclusively with the realization of one- and two-ports, solve both problems simultaneously through the concept of canonical structures: the given function is first decomposed according to a topological configuration particularly simple and involving the minimum number of elements; then it is proved that the necessary conditions for realizability, deduced from general physical considerations, lead to realizable network elements on such a structure in all cases.

The synthesis of single-element-kind n-ports is essentially an unsolved problem, since no canonical structure is known per $n \geq 4$*; moreover, the apparently evident question of their existence, that is, of the realizabillity of an n-port on a network including the minimum number of elements, which has been recently raised by G. Biorci,(2) is still open, except in some particular cases.(3)

On the other hand, even if one knew a network configuration, or a class of them which could be considered canonical for all or some n-ports, there would still be the serious difficulty of finding their element values from the knowledge of its terminal performance. Such a difficulty, intimately connected with the general problem of solving network functions with respect to the branch constants, probably calls for a better knowledge of the particular constitution of such functions.

The intricate state of the matter led me and others to the conviction that the possibility of enlightening the problem depends strictly on the success of a wide investigation of the topological and electrical properties of network functions; this investigation should lead to the knowledge of canonical structures, and perhaps it could give some information on the functional dependence of the branch constants from the network parameters, which appears to be very involved in a purely algebraic approach.

---

*For $n = 2$ one has to do with the familiar T and Pi realizations of two ports; the case $n = 3$ has been solved by Tellegen.(1)

The first step accomplished in this direction was the investigation of degenerate n-port structures, that is, of structures which force a functional dependence among the network parameters, i.e., a relation identically fulfilled independently from the branch values.

The interest of such research was evident from the fact that it was aimed at an identification of the canonical structures by systematically excluding those affected by some inner lack of generality.

The first results of this investigation were condensed in a paper,(4) in which we gave some properties of the so-called linear degeneracies and a few concepts, or more precisely a method of investigation, for the nonlinear ones. The complete theory of degenerate n-ports was left for further research. It became clear that the extreme complexity of the problem suggested coming back to the linear degeneracies in order to derive a complete theory of them and investigate their ultimate implications.

This is done in the present work, where it is proved that all linear degeneracies in n-ports simply express Kirchhoff's current and voltage laws. The network structures, which lead to such degeneracies, are investigated.

II. DEFINITIONS

Some definitions and general concepts are listed for the sake of clarity.

An n-port is an open network which interacts with n external one-ports through n terminal pairs. An important particular case is found when the n terminal pairs are tied together, so that by joining the nodes of each pair by an arrow, one obtains a tree: such an n-port, evidently based on a $(n + 1)$-node network, will be called an $(n + 1)$-pole.

An n-port structure is the linear graph of an n-port, that is, the figure which is obtained by substituting each branch by a line and disregarding the branch constants.

The order of an n-port structure is the number of ports which should be added to the given ones in order to obtain a tree. This number will be indicated by p . It is clear that any n-port can be thought of as an $(n + p + 1)$-pole, p ports of which are permanently open-circuited. The external performance of an n-port can be described either by its port admittance or by its port impedance matrix: the first representation is more convenient when n-port is fed by voltage sources, the second when it is fed by current generators.

An n-port structure will be said to be strictly degenerate if either its admittance matrix Y or its impedance matrix Z or both do not exist. These cases are so extreme that we can disregard them a priori: in fact, the degeneracy can be ascertained by simply glancing at the network diagram.

An n-port structure will be said degenerate if there is at least one relation which one or more of the elements in the Y or Z matrix must identically fulfull, independently of the branch constants. Such a definition shows that the degeneracy is a property of n-port structures, rather than of n-ports: it affects, not a particular network, but all networks having the same topological configuration.

Degeneracies can be conveniently classified, according to the algebraic form of the relations which express them, as linear and nonlinear degeneracies.

III. GENERAL THEOREMS

In this section, I simply report three general theorems, proved in the previously quoted paper,(4) which will be extensively used in the following:

Theorem 1: Let an n-port structure of order p be degenerate: then, any n-port structure of order p-1 obtained from it by considering any one of the additive ports as an actual one (provided that it is chosen in parallel with an actual branch) is affected by the same degeneracy.

Theorem 2: Any (n + 1)-pole structure based on a complete graph is nondegenerate.

By combining theorems 1 and 2, one easily reaches the following conclusion:

Theorem 3: Any n-port structure based on a complete graph is nondegenerate.

From theorem 3, it follows that any degeneracy equation must express, possibly in some intricate fashion, the fact that some branches lack in the actual structure. In the following, we shall study what configuration these lacking branches must form in order to cause a linear degeneracy.

IV. ANALYTICAL DESCRIPTION OF THE (N + 1)-POLE

The electrical performance of an (n + 1)-pole is usually described either on an admittance or on an impedance basis, i.e., by assuming either the n-port voltages or the n-port currents as independent variables. The first representation is far more convenient since the port admittances are linear functions of the branch constants, while this does not happen for the second one. Hence we shall first deal exclusively with the former. However, it is clear that any set of n variables, depending on the n-port voltages through a linear nonsingular transformation, can be used as well to accomplish our aim: the nonsingularity of the transformation ensures of their independence, while its linearity preserves the above said property.

It will be shown in the following treatment that linear degeneracies are described most conveniently in terms of such generalized network variables: hence I provide, in this chapter, the fundamental concepts on which this representation bears.

A cut-set of a linear graph is defined as a set of branches dividing all nodes into two classes, such that it is impossible to go from a node in the one class to a node in the other without passing through a branch in the cut-set.(5)

A cut-set surface is a closed surface including just nodes of either class of nodes defined by a cut-set and hence cutting just all branches of that cut-set. A cut-set surface as well as the branches of the graph may be arbitrarily oriented.

In the following, cut-sets will be denoted by Latin letters a and b while branches will be labelled with Greek letters λ and μ.

The cut-set matrix of a linear graph is a matrix which has its rows associated with the cut-sets and its columns with the branches of the graph. The entry at the overlap of a row a and column λ is ± 1, if branch λ belongs to the cut-set a and cuts its associated surface in the same or in the opposite sense, is 0 if branch does not belong to the cut-set a.

The rank of the cut-set matrix of an $(n + 1)$-node graph is just n. n cut-sets of an $(n + 1)$-node graph are said to form an independent set, if n rows pertinent to the cut-set matrix are removed. The sub-matrix built up from such rows and all columns has rank n. To any cut-set b one can associate an algebraic variable V_b, which will be called cut-set voltage for reasons which will become clear in the following. The very important point is that two any sets of independent cut-set voltages are related by a linear nonsingular transformation.

All these definitions and properties are extensively treated in the literature.(6,7)

Now, we can define formally the admittance matrix pertinent to some independent set of cut-sets by the following equation:

$$Y_{ab} = \sum_{\lambda} D_{a\lambda}\, y_{\lambda\lambda}\, D_{\lambda b} \qquad (1)$$

where $y_{\lambda\lambda}$ is the admittance of branch λ and coefficients $D_{a\lambda}$ and $D_{b\lambda}$ are the entries of the cut-set matrix at the overlaps of row a with column λ and of column λ with row b respectively.

The definitions involved in Equation (1) can be made explicit by the following statements:

Definition 1: the self-admittance of a cut-set a is the sum of the admittance of the branches cutting its surface, all taken as positive.*

Definition 2: the mutual admittance between any two cut-sets a and b is the sum of the admittances of the branches cutting both their surfaces, taken as positive if the surfaces are cut in the same direction, as negative otherwise.

These definitions are illustrated in Figure 1: the self-admittance of the cut-set b is

$$Y_{bb} = y_{11} + y_{22} + y_{44} + y_{66}$$

while the mutual admittance between cut-sets a and b is

$$Y_{ab} = y_{11} - y_{22}.$$

---

*It is obvious that the self-admittance of a cut-set a is obtained by putting b = a in Equation (1).

In order to endow these concepts with an electrical content, feed all ports cutting surface b[*] by unit voltage sources all of whose positive terminals lie on the positive side of the surface; let all other ports be short-circuited (Figure 2). Then it is clear that all nodes lying on the same side of b are the same potenial and that the potential of nodes on the positive side is 1 V greater than that of the nodes on the negative side. Hence all branches not cutting surface b are either materially (if they are in parallel with ports) or electrically short-circuited and and do not carry any current. On the contrary, branches cutting surface b are all subjected to unit voltages and therefore carry currents numerically equal to their admittances, directed from the positive to the negative side of the surface. For example, in Figure 2, branch 5 is materially short-circuited, whereas branch 3 is electrically short-circuited with its terminals at the same potential; current is carried through branches 1, 2, 4, 5 towards the negative side of surface b. But, according to definition 1, the sum of such currents, obviously all taken as positive, is just the self-admittance of the surface b.

Moreover, some branches cutting surface b also cut, in general, surface a (for example, branches 1 and 2 respectively, in Figure 2). Their currents, numerically equal to their admittances, go from the positive to the negative side of the latter if they cut both surfaces in the same direction, from the negative to the positive in the opposite case. But, according to definition 2, the algebraic sum of such currents (taken as positive in the first case, as negative in the second one) is just the mutual admittance between the surfaces a and b.

---

[*] In the following, surfaces associated with cut-sets a and b will be more briefly quoted as surfaces a and b.

Now, if one takes into account Kirchhoff's Current Law, the previous conclusions can be reinterpreted in terms of port rather than of branch currents; or, which is the same under the given excitation, in terms of port self and mutual admittances rather than of branch admittances.

The sum of all branch currents cutting a surface from its positive to its negative side is just equal to the sum of all port currents cutting the surface in the opposite direction: that is, of all port currents taken as positive or negative according as the ports and the surface are the same or the opposite direction. On the other hand, each port current caused by the said excitation, is just equal to the sum of the mutual admittances between itself and the excited ports, taken as positive for ports fed in the positive sense, as negative for ports fed in the negative one; obviously, if the port under consideration is itself excited, its self-admittance must be added with the same sign rule.

The conclusion is that, if we denote the ports by the indices u and v, the admittance matrix pertinent to any set of independent cut-sets can be expressed in terms of the port admittance matrix as

$$Y_{ab} = {}_{u}\Sigma_{v}\ D_{au}\ Y_{uv}\ D_{vb} \qquad (2)$$

where D_{au} is ± 1 or 0 according as port u cuts surface a in the same or in the opposite sense, or does not cut it. D_{vb} is analogously defined.

Actually, Equation (2) is nothing but a particular case of the transformation which an admittance matrix undergoes when it passes from one to another set of independent cut-sets: ports, indeed, as they are organized to form a tree, define a particular independent set (which is called fundamental), such that each surface of it cuts exactly one port.(6,7)

At this point, it becomes clear the advantage of defining a cut-set (or surface) voltage as the common voltage of a set of sources connect across its surface in the above fashion; for the sake of brievity, a surface cut by fed ports will be called an excited surface. On the other hand, it is useful to denote the sum of the port currents cutting a surface from its negative to its positive side as a cut-set (or surface) current.

Hence the mutual admittance between cut-sets a and b is concisely defined as the short-circuit surface current* of a when b is excited by unit voltage sources and all other independent surfaces (or ports) are short-circuited; an obviously similar definition holds for the self-admittance of a cut-set.

---

*Here a surface is considered as short-circuited, when all ports cutting it are short-circuited except the ones which are fed by sources because they belong also to an excited surface.

V. ANALYTICAL DESCRIPTION OF THE N-PORT

All the previous concepts were referred to the case of an (n + 1)-pole network. It is not difficult, however, to extend them to a general n-port.

In order to do this, it is better to state a systematic notation.

Let an n-port structure based on an (n + p + 1)-pole network be considered, and let the actual ports be numbered from 1 to n, the additive ones from $n+1$ to $n+p$. Then the n-port can be thought of as being obtained from the (n + p + 1)-pole by opening ports $n+1, \ldots, n+2, \ldots, n+r, \ldots, n+p$ in any order. Since numbering is arbitrary, there is no lack in generality if the ports are assumed to be opened in the same order as they are numbered. The (n + p + 1)-pole mutual admittances between any two ports u and v or any two surfaces a and b will be denoted by the symbols Y_{uv} and Y_{ab}, as in the preceding part of the paper; the (n + p - r)-port mutual admittances between ports u and v and surfaces a and b will be denoted as $Y^{(r)}_{ur}$ and $Y^{(r)}_{ab}$ respectively. The latter concept obviously requires defining. The r-th order* mutual admittance $Y^{(r)}_{ab}$ between two surfaces a and b is the surface current crossing a, when surface b is excited by unit voltage sources, and ports from 1 to n and from $n + r + 1$ to $n+p$ are short-circuited, while ports from $n+1$ to $n+r$ are open. A strictly similar interpretation holds for the self-admittance of a surface.

---

*In the following any parameter pertinent to a (n + p -r)-port will be called an r-th order parameter.

In order for the definition to be consistent, neither one of the surfaces a and b should cut any one of the open circuited ports.

It is well known that (n + 1)-th order port admittances can be derived from r-th order ones by the recurrent formula:

$$Y_{uv}^{(r+1)} = Y_{uv}^{(r)} - \frac{Y_{u,n+r+1}^{(r)} \; Y_{n+r+1,v}^{(r)}}{Y_{n+r+1,n+r+1}^{(r)}} \quad . \qquad (3)$$

A similar formula can be proved for any two surfaces fulfilling the above consistency condition. Let surface b be excited by a cut-set voltage V_b and port n+r+1 be fed by a current source I_{n+r+1}. Let all other ports from 1 to n and from n+r+2 to n+p be short-circuited, while ports from n+1 to n+r are open. Then, if I_a denotes the surface current at a, one can write the following equations:

$$I_a = Y_{ab}^{(r)} V_b + Y_{a,n+r+1}^{(r)} V_{n+r+1}$$

$$I_{n+r+1} = Y_{n+r+1,b}^{(r)} V_b + Y_{n+r+1,n+r+1}^{(r)} V_{n+r+1} \; .$$

Opening port n+r+1 amounts to equating the left side of the second equation to zero. If V_{n+r+1} is then eliminated one finds:

$$Y_{ab}^{(r+1)} = Y_{ab}^{(r)} - \frac{Y_{a,n+r+1}^{(r)} \; Y_{n+r+1,b}^{(r)}}{Y_{n+r+1,n+r+1}^{(r)}} \; .$$

It is now easy to prove that, if relation (2) holds between the mutual admittance of any two surfaces a and b (or the self-admittance of a surface a) and the port admittances of an (n+p+1)pole network, it still holds when ports from n+1 to n+r are permanently opened, if the consistency condition is fulfilled.

Let us assume that Equation (2) holds for (r+1)-th order parameters. Then:

$$Y^{(r+1)}_{ab} = \sum_u \sum_v D_{au} Y^{(r-1)}_{uv} D_{vb} .$$

Now $Y^{(r)}_{ab}$ can be expressed in terms of $Y^{(r-1)}_{ab}$ by means of Equation (4), to yield:

$$Y^{(r)}_{ab} = \sum_u \sum_v D_{au} Y^{(r-1)}_{uv} D_{vb} - \frac{1}{Y^{(r-1)}_{n+r,n+r}} \sum_u D_{au} Y^{(r-1)}_{u,n+r} \sum_v Y^{(r)}_{n+r,v} D_{vb} . \qquad (5)$$

By collecting terms in the right side of Equation (5) one obtains

$$Y^{(r)}_{ab} = \sum_u \sum_v D_{au} \left(Y^{(r-1)}_{uv} - \frac{1}{Y^{(r-1)}_{n+r,n+r}} Y^{(r-1)}_{u,n+r} Y^{r-1)}_{n+r,v}\right) D_{vb} .$$

The term within parenthesis is nothing but $Y^{(r)}_{uv}$ according to Equation (4). Hence:

$$Y^{(r)}_{ab} = \sum_u \sum_v D_{au} Y^{(r)}_{uv} D_{vb} \qquad (6)$$

which proves the validity of Equation (2) for the parameters of order r . Since it holds for $r=o$, the theorem is proved for any r .

Equation (6) is simply a mathematical formulation of the physical concepts explained above.

VI. THE DEGENERATE (N+1)-POLE

The elements of the port admittance matrix of an (n+1)-pole are equal in number to the branches in the complete graph with (n+1)-nodes: this number, obviously equal to $n\ (n+1)/2$ will be denoted as l.

From theorem 2 we know that any complete (n+1)-pole is nondegenerate: on the other hand, if any s branches are missing from an actual structure, we have l port parameters depending on $l-s$ branch constants only. By eliminating the latter in the equations which express the port parameters as functions of such constants, one finds s relations identically fulfilled by the port parameters, i.e., s equations of degeneracy.

In order to derive them explicitly, we only need to remark that the admittance y_{ab} of the branch connecting any two nodes a and b is just the mutual admittance, with the sign reversed, between two surfaces drawn around nodes a and b respectively, both oriented either towards or away from them; hence, according to Equation (2), y_{ab} is expressed as:

$$- y_{ab} = \sum_u \sum_v D_{au}\ Y_{uv}\ D_{vb}$$

where D_{au} and D_{vb} are defined as in connection with Equation (2).

Therefore we may conclude that the lack of the branch connecting any two nodes a and b lead to the degeneracy equation:

$$\sum_u \sum_v D_{au}\ Y_{uv}\ D_{vb} = 0 \qquad (7)$$

Its physical meaning is evidenced along the general lines described in Chapter 4.

For example, in the network of Figure 3, the lack of the branch connecting nodes a and b leads to the equation:

$$Y_{17} - Y_{18} + Y_{19} - Y_{27} + Y_{28} - Y_{29} + Y_{37} - Y_{38} + Y_{39} = 0.$$

The result of Equation (7) must be carefully analyzed, since it is the basis for successive steps.

Each degeneracy equation pertinent to an (n+1)-pole either has the form (7) or it is a consequency of equations of such a form.

Therefore, if such equations are called fundamental, it follows that an (n+1)-pole is affected by as many fundamental degeneracies as there are lacking branches: all the pertinent equations are linear and of the rectangular form (7), they fulfill the sign rule implied by the definition of the coefficients D_{au} and D_{vb} and express Kirchhoff's Current Law at either one of the end nodes of the lacking branch when the surface associated with the other is excited.

At a first sight, the study of dependent degeneracies could appear fully uninteresting: but, as we shall see, this is not the case, since it often happens that a dependent degeneracy of an (n+p+1)-pole is not cancelled when some p of its ports are permanently open circuited. Hence the network is used as an n-port, while the fundamental degeneracies from which that one was derived no longer appear in the new situation.

Of all possible dependent degeneracies, only the linear ones will be investigated in this paper. Moreover, it will be shown that the class of the physically meaningful linear degeneracies allows a very simple electrical interpretation.[*]

VII. KIRCHHOFFIAN DEGENERACIES

An n-port structure of the r-th order will be said to be affected by a Kirchhoffian degeneracy, if one can draw upon it two closed surfaces a and b , such that, if either one of them is excited, the surface current through the other is identically equal to zero.

Mathematically, this definition can be concisely expressed in either one of the forms:

$$Y_{ab}^{(r)} = o \qquad (8)$$

$$\sum_u \sum_v D_{au} Y_{uv}^{(r)} D_{vb} = o \qquad (8')$$

which state the fact that the mutual admittance between surfaces a and b is zero.

It is interesting to note that surfaces a and b play symmetrical roles in the definition, since their exchange does not alter their mutual admittance by the reciprocity law, which is fulfilled by definition.(6)

---

[*] The analysis of the degeneracies arising from the dual condition $z_{ab} = o$ is not carried out since they merely reduce the number of ports (if the short-circuited branch is in parallel with a port), or lead to strict degeneracies which can be observed at sight: in fact, short-circuiting a branch not in parallel to a port amounts to forming a mesh including only ports; hence the port voltages are no longer independent and the Y matrix looses its existence. The same happens to the Z matrix only if all branches cutting a closed surface are opened.

In the rest of this section and in the next one, we shall deal with the problem of investigating the properties of the network structures, which are affected by Kirchhoffian degeneracies.

First, we prove the following.

Theorem 4: A necessary and sufficient condition in order that the mutual admittance $Y_{ab}^{(r)}$ between two surfaces a and b drawn upon an n-port structure be zero, is that surfaces a and b and all additive ports can be partitioned into three classes, say A, B, and C, such that A includes surface a; B includes surface b; any surface in A or in B has non-zero mutual admittance of order zero with at least another surface in the same class; and any surface in A, B, or C has zero mutual admittance of order zero with all surfaces of the other classes.

In particular, class A many contain only surface a, or class B only surface b: all additive ports are then divided between the other two classes, say B and C in the first case, A and C in the second one. Finally, class C may be empty.

Before proving this theorem, it is convenient to describe its content pictorially, as in Figure 4. Here points represent surfaces and each line represents a non-zero mutual admittance of order zero between the surfaces represented by its end points.

The proof of the theorem is carried out by mathematical induction from order r to order r+1. Suppose it is true for order r.

A necessary and sufficient condition in order that the mutual admittance $Y_{ab}^{(r+1)}$ be equal to zero is that the following conditions are fulfilled:

a) The mutual admittance $Y_{ab}^{(r)}$ is equal to zero.

b) Either one or possibly both the mutual admittances $Y_{a,n+r+1}^{(r)}$ and $Y_{n+r+1,b}^{(r)}$ are equal to zero.

This follows from Equation (3) and theorem 1.

Condition a) implies that surfaces a and b and all additive surfaces from n+1 to n+r can be divided into the three classes A, B, and C, since the theorem is true for order r.

Suppose now that condition b) is fulfilled because $Y_{a,n+r+1}^{(r)} = o$. Then, since the theorem is true for order r, surface n+r+1 cannot belong to class A and hence it must belong to either class B or C or possibly to both.* The same reasoning applies if $Y_{b,n+r+1}^{(r)} = o$, by simply exchanging class A with class B.

If both equations are fulfilled, then surface n+r+1 can belong neither to class A nor to class B and hence it belongs to class C.

This means that a necessary and sufficient condition in order that $Y_{ab}^{(r+1)}$ be equal to zero is that, if the conditions of the theorem are already fulfilled for all additive ports from n+1 to n+r, then port n+r+1 falls in either one of the classes, hence fulfilling such conditions.

---

*This case happens when port n+r+1 has non-zero mutual admittance of order r with at least one port in class B and at least one port in class C: then classes B and C collapse in an unique class if the second one is such that it is possible to pass from any port to another one through a continuous path of non-zero admittances; otherwise class C looses some ports which pass into class B.

We can therefore conclude that, if the theorem is true for order r, it remains true for order r+1.

It remains to be proved for $r=2$[*]. By applying Equation (4)twice and expressing all terms in its right side as functions of the zero th-order admittances one obtains:[**]

$$Y_{ab}^{(2)} = Y_{ab} - \frac{Y_{n+1,n+1}}{\Delta} Y_{a,n+2} Y_{n+2,b} - \frac{Y_{n+2,n+2}}{\Delta} Y_{a,n+1} Y_{n+1,b} + \frac{Y_{n+1,n+2}}{\Delta} (Y_{a,n+1} Y_{n+2,b} + Y_{a,n+2} Y_{n+1,b}) \quad (9)$$

where

$$\Delta = \begin{vmatrix} Y_{n+1,n+1} & Y_{n+1,n+2} \\ Y_{n+2,n+1} & Y_{n+2,n+2} \end{vmatrix}$$

The left side of Equation (9) is equal to zero, if and only if $Y_{ab} = 0$ and the sum of the other three terms in the right side is equal to zero (theorem 1). But such a sum vanishes if and only if each of its three terms is equal to zero: in fact, the numerator of the first term contains the admittance of the branch in parallel with port n+1, which does not appear in the other one. The numerator of the second one contains the admittance of the branch in parallel with port n+2, which fulfills the same condition: such two terms cannot compensate for each another or for the third one; hence each one of them must be equal to zero, and so must the last one.

---

[*] There is no interest in proving the theorem for $r=o$ or $r=1$ since its conditions appear only partially in such cases.

[**] The following development could be carried out more easily by using Baranov's theorem.(8)

Now we have to investigate the conditions under which such a situation arises. The first term is equal to zero if either one of the following equations holds:

$$Y_{a,n+2} = 0 \qquad Y_{n+2,b} = 0 \tag{10}$$

while the second term vanishes according as either one of the equations is fulfilled

$$Y_{a,n+1} = 0 \qquad Y_{n+1,b} = 0 \; . \tag{11}$$

Hence the following four cases are possible:

(a) $Y_{a,n+2} = 0 \; ; \; Y_{a,n+1} = 0$

(b) $Y_{a,n+2} = 0 \; ; \; Y_{n+1,b} = 0$

(c) $Y_{a,n+1} = 0 \; ; \; Y_{n+2,b} = 0$

(d) $Y_{n+1,b} = 0 \; ; \; Y_{n+2,b} = 0$.

In cases (a) and (d) the third term disappears. This is the case in which additive ports n+1 and n+2 both fall in either class B (case a) or in class A (case d), while class C is empty.

In cases (b) and (c) the third term vanishes if and only if $Y_{n+1,n+2}$ is equal to zero. This is the case in which port n+1 falls in class A and port n+2 in class B (case b) or vice versa (case c) while class C is empty; then $Y_{n+1,n+2} = 0$ as required by the theorem.

Finally, if equations of case (b) and the second one of case (d) are fulfilled, port n+1 belongs to class A and port n+2 to class C; it is easy to see that any combination of the equations pertinent to one of the cases a, b, c, d, with one or both the equations pertinent to some others simply yield different redistributions of the additive ports among the classes A, B, and C. For r=1, the additive port n+1 must have vanishing zero-th order mutual admittance with at least one of the surfaces a and b. Accordingly it must be an element of either one of the previously said classes.

Theorem 4 is very important since it allows a simple process for building up n-port structures affected by Kirchhoffian degeneracies, by simply stating that any two surfaces have vanishing zero-th order mutual admittance if and only if no branch exists which cuts both of them. This follows from Equation (1), where the branch admittances are regarded as independent variables, and where each non-zero $y_{\lambda\lambda}$ must consequently correspond to a vanishing $D_{a\lambda}$ or $D_{b\lambda}$ or possibly both.

In figure 5 an example of such a process is given for a first-order n-port structure. It shows several interesting features: first, there is no node having vanishing zero-th order mutual admittance with the additive port 9; hence it is proven that a dependent degeneracy on an (n+p+1)-pole network can become a fundamental one on the derived n-port. Moreover, it can be verified that the experimental detection of the degeneracy calls for a simultaneous feeding of several ports. It is impossible to simplify either surface a or b in such a way that it cuts one port only; this fact is due to the necessity of short-circuiting electrically some branches which otherwise would be traversed by currents. It must be remarked that no transformation of the reference frame allows such a simplification, since the surfaces intersect mutually and hence do not correspond to the ports of any tree.

VIII. APPLICATION OF THE GENERALIZED COHN'S THEOREM TO THE ANALYSIS OF KIRCHHOFFIAN DEGENERACIES

In order to give a better insight into the properties of Kirchhoffian degeneracies, a more general distribution of sources on the n-port structure must be considered. Suppose that, in a network with $n+p$ ports and b branches, each port v is fed by a voltage source V_v connected across its terminals and each branch μ is fed by a voltage source E_μ inserted in series.

Let $Y_{u\mu}$ denote the zero-th order mutual admittance between port u and branch μ, that is, the current flowing into port u when branch μ is fed by a unit source in the positive sense, and all other voltage sources are short-circuited; let $Y_{\lambda\mu}$ represent the zero-th order mutual admittance between branches λ and μ, for which a similar interpretation holds.

The superposition principle allows the following set of equations for the (n+p+1)-pole network:

$$\begin{aligned} I_u &= \sum_{v=1}^{n+p} Y_{uv} V_v + \sum_{\mu=1}^{b} Y_{u\mu} E_\mu \qquad \begin{matrix} (u=1,2,\ldots,\ n+p) \\ (\lambda=1,2,\ldots,\ b) \end{matrix} \\ I_\lambda &= \sum_{v=1}^{n+p} Y_{\lambda v} V_v + \sum_{\mu=1}^{b} Y_{\lambda\mu} E_\mu \ . \end{aligned} \tag{12}$$

The detailed analysis of the set of Equations (12) and of its connection with the generalized Cohn's Theorem is done in another paper.(9) Here I describe the properties of main interest to our aim.

a) The admittance parameters of Equations (12) are related to the branch admittances by the following equations:

$$\begin{aligned} Y_{uv} &= \sum_{\lambda} D_{u\lambda}\, y_{\lambda\lambda}\, D_{\lambda v} \\ Y_{u\mu} &= D_{u\mu}\, y_{\mu\mu} \\ Y_{\lambda v} &= y_{\lambda\lambda}\, D_{\lambda v} \\ Y_{\lambda\mu} &= \begin{cases} y_{\lambda\lambda} & \text{if } \mu = \lambda \\ 0 & \text{if } \mu \neq \lambda \end{cases} \end{aligned} \tag{13}$$

From inspection of Equations (13), it appears that the matrix of Equations (12) is symmetrical, so that the newly introduced parameters fulfill the reciprocity law.

b) The network description in terms of Equations (12) is highly redundant since the following relationships hold:

$$\begin{aligned} Y_{uv} &= \sum_{\lambda} D_{u\lambda}\, Y_{\lambda v} \\ Y_{uv} &= \sum_{\mu} Y_{u\mu}\, D_{\mu v} \\ Y_{uv} &= \sum_{\lambda}\sum_{\mu} D_{u\lambda}\, Y_{\lambda\mu}\, D_{\mu v} \ . \end{aligned}$$

These relationships show that the matrix of Equations (12), which will be called the enlarged admittance matrix of the (n+p+1)-pole, has exactly rank b.

c) Suppose now that ports n+1, n+2,..., n+r on the (n+p+1)-pole network are left permanently open. The performance is now described by a set of equations like (12) where zero-th order parameters are replaced by r-th order ones and sums from 1 to n+p are replaced by sums ranging over ports from 1 to n and from n+r+1 to n+p. The problem arises of giving reduction formulae relating the r-th order with the (r+1)-th order parameters. For the mixed and the branch-to-branch admittances, one finds:

$$Y_{u\lambda}^{(r+1)} = Y_{u\lambda}^{(r)} - \frac{Y_{u,n+r+1}^{(r)}\ Y_{n+r+1,\lambda}^{(r)}}{Y_{n+r+1,n+r+1}^{(r)}} \qquad (14)$$

$$Y_{\lambda\mu}^{(r+1)} = Y_{\lambda\mu}^{(r)} - \frac{Y_{\lambda,n+r+1}^{(r)}\ Y_{n+r+1,\mu}^{(r)}}{Y_{n+r+1,n+r+1}^{(r)}} \ . \qquad (15)$$

The generalized Cohn's theorem expresses the derivatives of the port r-th order admittances in terms of the mixed parameters in the following way:(8)

$$\frac{\partial\, Y_{uv}^{(r)}}{\partial Z_{\lambda\lambda}} = -Y_{u\lambda}^{(r)}\ Y_{\lambda v}^{(r)} \ . \qquad (16)$$

Equation (16) leads to the following:

Theorem 5: A necessary and sufficient condition in order that an r-th order n-port structure be affected by a Kirchhoffian degeneracy is that one can draw on it two closed surfaces a and b such that every branch has vanishing r-th order mutual admittance with at least one of them.

Let a Kirchhoffian degeneracy described by equation

(8') be considered. Let the impedance of a branch λ be increased by an infinitesimal amount $dz_{\lambda\lambda}$. Since equation (8') is identically fulfilled, independently of the branch parameter values, we obtain:

$$\sum_u \sum_v D_{au} \left(Y^{(r)}_{uv} + \frac{\partial Y^{(r)}_{uv}}{\partial Z_{\lambda\lambda}} dZ_{\lambda\lambda} \right) D_{vb} = 0. \qquad (17)$$

By subtracting Equation (8') from (17) we obtain:

$$\sum_u \sum_v D_{au} \frac{\partial Y^{(r)}_{uv}}{\partial Z_{\lambda\lambda}} D_{vb} = 0 \qquad (18)$$

for any λ.

If now one takes into account Equation (16) by substituting its right side into Equation (18), the following result is achieved:

$$\sum_u \sum_v D_{au} Y^{(r)}_{u\lambda} Y^{(r)}_{\lambda v} D_{vb} = 0.$$

That is:

$$\sum_u D_{au} Y^{(r)}_{u\lambda} \sum_v D_{bv} Y^{(r)}_{v\lambda} = 0. \qquad (19)$$

By a straightforward generalization of the previously stated notation, any branch λ either $Y^{(r)}_{a\lambda}$ or $Y^{(r)}_{b\lambda}$ or possibly both are equal to zero, which is precisely the statement of the theorem.

One could inquire under which topological conditions either one of these equations holds. By a reasoning perfectly parallel to the one carried out in proving theorem 4, one finds that $Y^{(r)}_{a\lambda}$ vanishes if and only if branch λ can be included either in class B or in class C: that is, if branch λ has non-zero mutual admittance only with those additive ports, each of which has a vanishing mutual admittance with surface a. Similarly for $Y_{b\lambda}^{(r)} = 0$. Hence classes A, B, and C can be redefined to include branches, and result in the following:

Theorem 6: A necessary and sufficient condition in order that an n-port structure be affected by a Kirchhoffian degeneracy, is that any branch can be included in either one of the previously defined classes A, B, and C, with the convention that a branch included in a class has vanishing mutual admittance with the ports and the surfaces belonging to the others.

Up to this point, we have stated that, if $Y_{ab}^{(r)} = 0$, the n-port structure is degenerate since the surface current of a is identically zero when surface b is excited and all other actual ports are short-circuited. One can inquire whether the currents in branches cutting surface a are all equal to zero, or, according to Kirchhoff's Current Law, only their sum vanishes. In the case of Figure 5 the first alternative was true, but the property holds in general, as stated from the following:

Theorem 7: Consider an r-th order n-port structure on which two surfaces a and b have been drawn such that $Y_{ab}^{(r)} = 0$. Then the fact that the surface current of a, when surface b is excited, is identically equal to zero expresses Kirchhoff's Current Law for such a surface, since the current in each branch cutting it is identically equal to zero.

Let b be the excited surface. Let us consider a branch λ cutting surface a.

Since $Y^{(r)}_{ab} = 0$, by theorem 6, either $Y^{(r)}_{a\lambda}$ or $Y^{(r)}_{\lambda b}$ or possibly both must vanish. The theorem is proven if one is able to show that $Y^{(r)}_{\lambda b}$ is zero in all cases, since this means that no current flows in branch λ when surface b is excited. It is then sufficient to prove that $Y^{(r)}_{a\lambda}$ cannot be zero. In fact, if $Y^{(r)}_{a\lambda}$ is zero, so $Y_{a\lambda}$ would also be.* But this is in conflict with the hypothesis that branch λ cuts surface a. In fact, let all actual and additive ports be short-circuited, and let branch λ be series fed by a unit voltage source; then all branch currents but I_λ are zero since all branches are short-circuited: I_λ in turn, is numerically equal to $y_{\lambda\lambda}$. Such a current crosses surface a: hence there must be, across a, some other currents which together with the latter give a sum equal to zero in such a way as to fulfill Kirchhoff's Current Law: the latter are necessarily the currents of the ports cutting a, since all branches are short-circuited. But this means precisely that $Y_{a\lambda} = \pm y_{\lambda\lambda}$, which is different from zero.

This theorem can also be proven along a quite different line. Branches cutting surface a either belong to some additive cut-set, or not. In the latter case, they are either electrically or materially short-circuited and hence they do not carry any current; in the first one, let λ be the branch taken into consideration and let an additive port, say $n+r+1$, be placed across it. Then, by Equation (4), $Y_{ab}^{(r+1)} = 0$ implies $Y_{ab}^{(r)} = 0$ and moreover that either $Y_{a,n+r+1}^{(r)}$ or $Y_{n+r+1,b}^{(r)}$ or both are equal to zero. But $Y_{a,n+r+1}^{(r)}$ cannot vanish since this would imply $Y_{a,n+r+1}^{(r)} = 0$, which is false, since both surfaces a and $n+r+1$ cut at least branch λ. Hence $Y_{n+r+1,b}^{(r)}$ must be equal to zero. This means that, if surface b only is excited, with all ports not cutting it from 1 to n and from $n+r+1$ to $n+p$ short-circuited, no current flows in port $n+r+1$. This means that, if the latter is opened, no voltage appears across it and hence no current flows in the branch connected across its end nodes. The reasoning is then repeated for all branches in the proper conditions and the theorem is proved.

---

*This can be easily shown following the same reasoning which lead to the proof of theorem 1, if branch λ is not connected across an additive port. If this is the case, one can repeat the reasoning by simply shifting the port in parallel with another branch: this operation obviously does not alter the $Y^{(r)}_{a\lambda}$'s and hence the conclusions about them are still valid. The only exception occurs when this is not possible since the considered branch is unique: this is the case of a hinged network, for which the proof of the theorem is immediate.

Electrically, all branches can be divided into two classes: those which do not belong to any additive cut-set, which are materially or electrically short-circuited; and those which belong to one or more additive cut-sets, which have the interesting property that in any mesh including them the impressed voltage sum is identically equal to zero. An example is given in Figure 5: branches 1 and 2 fall in the first class, where branches 3 and 4 fall in the second one.

Obviously there are many trivial particular cases: from the simplest one, which happens when surface a is cut by no branches at all, to the more refined one which arises when under the given excitation the network becomes hinged, so that no mesh exists, including the branches of the second class. These cases were already investigated in the previously quoted work.(4)

For the sake of completeness, we report here a corollary, which resumes the necessary and sufficient conditions for the existence of a Kirchhoffian degeneracy in terms of voltage and current ratios. It follows:

Corollary 1: Necessary and sufficient conditions for an r-th order n-port structure to be affected by a Kirchhoffian degeneracy associated with two surfaces a and b, are that the voltage across any branch is zero when either one of the said surfaces are excited by voltage sources; or that the current sum through either one of the two surfaces is zero when any branch is fed by a current source connected across its end nodes.

A last important result which follows from Cohn's theorem is that if the network does not become hinged when either one of the two considered surfaces is excited, then one of them has zero mutual admittance with all branches of the second class and with all additive ports. This is easily proved: let λ and μ be any two branches of the second class and let the equation:

$$Y_{a\lambda}^{(p)} = 0$$

be fulfilled.

Then also the equation

$$\frac{\partial Y^{(p)}_{a\lambda}}{\partial Z_{\mu\mu}} = 0$$

holds, and this can be written, by Cohn's generalized theorem, as

$$Y^{(p)}_{a\lambda} \; Y^{(p)}_{\mu\lambda} = 0.$$

Now let all ports in the n-port structure be short-circuited: then a network is obtained, where all branches are of the second class. The mutual admittance between any two of them, say λ and μ is non-zero provided there is at least a mesh including both of them; that is, the n-port structure with all ports short-circuited is not hinged. Hence $Y^{(p)}_{a\mu} = 0$ as stated above.

The second part of the theorem is proved by noticing that if r is any additive port, then:(9)

$$Y^{(p)}_{ar} = \Sigma \, Y^{(r)}_{a\lambda} \, Z_{\lambda\lambda} \, Y^{(p)}_{\lambda r} .$$

Note that $Y^{(p)}_{\lambda r}$ is different from zero only as λ is a branch of the second class; on the other hand, if $Y^{(p)}_{a\lambda} = 0$ for some value of λ it is zero for all, according to the first statement. Hence $Y^{(p)}_{ar} = 0$ for any value of r; that is, for all additive ports.

Hence we can state the following:

Corollary 2: If an n-port structure is such that it does not become hinged when all ports are short-circuited then a necessary and sufficient condition for a Kirchhoffian degeneracy is that there exist on it two surfaces a and b whose zero-th order mutual admittance vanishes and at least one of which has vanishing zero-th order mutual admittance with all additive ports.

The analysis of Kirchhoffian degeneracies is now complete; in the following chapter we prove a necessity theorem, which allows reduction of any linear degeneracy to a single Kirchhoffian degeneracy or to a linear sum of them.

IX. THE NECESSITY THEOREM

First, let us recall Equation (3):

$$Y_{uv}^{(r+1)} = Y_{uv}^{(r)} - \frac{Y_{u,n+r+1}^{(r)} \; Y_{n+r+1,v}^{(r)}}{Y_{n+r+1,n+r+1}^{(r)}} \qquad (3)$$

For the further developments, it is convenient to introduce a short-hand notation by putting

$$\frac{Y_{u,n+r+1}^{(r)}}{\sqrt{Y_{n+r+1,n+r+1}^{(r)}}} = h_u^{(r)} \qquad \frac{Y_{n+r+1,v}^{(r)}}{\sqrt{Y_{n+r+1,n+r+1}^{(r)}}} = h_v^{(r)} . \qquad (20)$$

Then the complete set of equations connecting (r+1)-th order to r-th order admittances on a given structure becomes:

$$\left\{ \begin{aligned} Y_{uv}^{(r+1)} &= Y_{uv}^{(r)} - h_u^{(r)} \; h_v^{(r)} \\ 0 &= Y_{u,n+r+1}^{(r)} - h_u^{(r)} \; h_{n+r+1}^{(r)} \\ 0 &= Y_{n+r+1,n+r+1} - (h_{n+r+1}^{(r)})^2 \quad . \end{aligned} \right. \qquad (21)$$

In a purely formal approach, the $h^{(r)}$'s can be regarded as algebraic variables, *a priori* independent, whose values are possibly constrained solely by the set of Equations (21).

Any linear degeneracy affecting the (r+1)-th order structure can be put in the form:

$$\sum_{uv} D_{uv} Y_{uv}^{(r+1)} = 0. \qquad (22)$$

By theorem 1 and according to the first of Equations (21) it follows that both of the following Equations must hold:

$$\sum_{uv} D_{uv} Y_{uv}^{(r)} = 0 \qquad (23)$$

$$\sum_{uv} D_{uv} h_u^{(r)} h_v^{(r)} = 0 . \qquad (24)$$

Now, it must be asked in what manner an equation such as (24) can arise from Equations (21).

Suppose the structural properties of the r-th order network are entirely known; then, by giving the $h^{(r)}$'s all values consistent with Equations (21), one should be able to deduce, from the same equations, the relationships among the $Y_{uv}^{(r+1)}$'s, and hence the structural properties of the (r+1)-th order network.

As long as the latter have not been deduced, the $Y_{uv}^{(r+1)}$'s are a priori independent parameters: hence no functional relationship among the $h^{(r)}$'s can be deduced from the first group of euqations in the set (21). Consequently, in order to investigate the possibility and the structure of such an equation as (24), we must limit our investigation to the second group and to the last equation in the above set.

First, let us prove the following:

Lemma: Any functional relationship among r-th order mutual admittances of a single additive port versus some actual ports is Kirchhoffian.

The proof is carried out by induction. That the lemma is true for $r = 0$ is evident, since we know that on an (n+p+1)-pole structure all degeneracies are Kirchhoffian or can be reduced to them.

Suppose now it is true for some $r-1$ and let us proove it remains valid for r. Let s be the additive port which is supposed to be short-circuited on both (r-1)-th and r-th order structures: Equation (3) yields:

$$Y^{(r)}_{us} = Y^{(r-1)}_{us} - \frac{Y^{(r-1)}_{u,n+r}\, Y^{(r-1)}_{n+r,s}}{Y^{(r-1)}_{n+r,n+r}} \quad . \tag{25}$$

By theorem 1 any functional relationship among the $Y^{(r)}_{us}$ (s fixed) must also be fulfilled by the $Y^{(r-1)}_{us}$. But the latter are assumed to be all of the form

$$\sum_u D_{au}\, Y^{(r-1)}_{us} = 0 \tag{26}$$

so that any functional relationship among the $Y^{(r-1)}_{us}$ must arise from the equations obtained by substituting (26) into (25).

$$\sum_u D_{au}\, Y^{(r)}_{us} = - \frac{Y^{(r-1)}_{n+r,s}}{Y^{(r-1)}_{n+r,n+r}} \sum_u D_{au}\, Y^{(r-1)}_{u,n+r} \tag{27}$$

Let Equation (27) be written as:

$$Y^{(r)}_{as} = - \frac{Y^{(r-1)}_{n+r,s}}{Y^{(r-1)}_{n+r,n+r}}\, Y^{(r-1)}_{a,n+r} \quad . \tag{28}$$

Any functional relationship identically fulfilled by the left sides of Equations (28) is clearly a homogeneous one, of some degree m, in such variables. Hence, by Euler's theorem, the same relationship must be fulfilled by the second factors of the right sides.

But, if the lemma is true for $r-1$, any relationship among the $Y^{(r-1)}_{a,n+r}$'s can be reduced to the form:

$$\sum_u D_{au} \, Y^{(r-1)}_{u,n+r} = 0$$

so that any relationship among the $Y^{(r)}_{as}$'s can be reduced to the form

$$\sum_u D_{au} \, Y^{(r)}_{us} = 0 \tag{29}$$

which proves the lemma.

Let us apply this result to the analysis of the functional relationships among parameters $h^{(r)}$'s . Let index s, in Equation (29), be assumed to have the value $n+r+1$. It is to be remarked that such a value cannot be assumed by index u, since the r-th order self admittance of the additive port $n+r+1$ is surely independent, owing to the branch connected across its terminals.

By taking into account Equations (21), we can state that any functional relationship among the $h^{(r)}$'s can be reduced to the form, equivalent to Equation (29):

$$h^{(r)}_{n+r+1} \sum_u D_{au} \, h^{(r)}_u = 0.$$

Since $h^{(r)}_{n+r+1}$ cannot be zero, because otherwise $Y^{(r)}_{n+r+1,n+r+1}$ would vanish and the network would be disconnected, it must be:

$$\sum_u D_{au} \, h^{(r)}_u = 0 . \tag{30}$$

The proof of the theorem is carried out by induction now. First it is time for $r=0$, as we have seen in Chapter 6. Let us suppose it is true for some r and prove it remains such for $r+1$.

In order that Equation (22) is fulfilled, both Equations (23) and (24) must be verified.

But, owing the above lemma, Equation (24) can be fulfilled if and only if one or more equations such as (30) are, so that it must split in the form:

$$\sum_{ab} D_{ab} \sum_{u} D_{au} h_{u}^{(r)} \sum_{v} D_{bv} h_{v}^{(r)} = 0 \tag{31}$$

where in every addend denoted by a pair of indices a and b, either one of which or possibly both, vanish.

On the other hand, Equation (23) splits similarly in the form:

$$\sum_{ab} D_{ab} \sum_{u} \sum_{v} D_{au} Y_{uv}^{(r)} D_{vb} = 0.$$

Such an equation must necessarily follow as a linear sum of the equations expressing the Kirchhoffian degeneracies which affect the r-th order structure. If surface pairs are denoted by the indices r and s, these equations can be written in the form

$$Y_{rs}^{(r)} = 0. \tag{32}$$

On the other hand, Equation (31) can be rewritten as

$$\sum_{ab} D_{ab} Y_{ab}^{(r)} = 0$$

where, by the above stated concept, each $Y_{ab}^{(r)}$ can be expressed in terms of the $Y_{rs}^{(r)}$'s:

$$Y_{ab}^{(r)} = \sum_{r} \sum_{s} D_{ar} Y_{rs}^{(r)} D_{sb} \tag{33}$$

and hence is zero. But this means that Equation (31) vanishes term by term, so that from the equations

$$Y_{ab}^{(r)} = 0 \tag{34}$$

and

$$h_a^{(r)} \; h_b^{(r)} = 0 \tag{35}$$

it follows

$$Y_{ab}^{(+1)} = 0. \tag{36}$$

It remains to prove that Equation (26) is truly Kirchhoffian; that is, that a and b are two surfaces on the network.

There is no lack in generality in supposing that vanishing terms in Equation (31) are those denoted by the index a. From the above lemma, it follows at once that a is a surface. As we are concerned with B, we must consider two cases: either coefficients D_{sb} in Equation (33) are topologically meaningful, so that b is also a surface, and then the question is solved: or coefficients D_{sb} do not have this property, and then Equation (35) can be further split into a sum of terms of the type:

$$Y_{as}^{(r+1)} = 0$$

where both a and s are surfaces. That such terms are zero is easily seen by simply noticing that Equations (34) and (35) can be split in the form:

$$\sum_s Y_{as}^{(r)} \; D_{sb} = 0$$

$$h_a^{(r)} \sum_s D_{bs} \; h_s^{(r)} = 0$$

which are zero term by term: the first one, owing to the fact that every $Y_{as}^{(r)}$ is equal to zero [Equation (32)] the second one owing to the fact that $h_a^{(r)}$ is zero. Hence the theorem is fully proven.

X. CONCLUSION

The essential results of the paper can be summarized as follows: first, Kirchhoffian degeneracies have been defined, and their electrical and structural peculiarities have been fully investigated; then, it has been proved that any linear degeneracy either is Kirchhoffian or can be resolved into a linear sum of Kirchhoffian ones. The still needed work on degeneracies concerns two main points: a) in order to give the previous theory all its practical value, a diagnosis technique should be devised to discover Kirchhoffian degeneracies on a given network structure without examining the mutual admittances among all possible cut-set pairs; b) a general theory of nonlinear degeneracies, so far fully investigated only for an n-port structure based on n+2 nodes (4), should be provided. As pointed out in the Introduction, the final goal of such investigation is a systematic classification of canonical structures, obtained by discarding the degenerate ones; this classification is crucial in order to prove Biorci's conjecture and its possible application to the synthesis of n-ports.

XI. REFERENCES

1. B.D.H. Telleger, *Theorie der Wisselstronen*, Groningen 1952, pp. 166-168.

2. G. Biorci, In "The realization of n-Port Networks without Transformers - A Panel Discussion," *IRE Trans. on Circuit Theory*, Vol. CT-9, (Sept. 1962), pp. 203-212.

3. G. Biorci, P. P. Civalleri, "About a Basic Theorem on Resistive Networks," *IEEE International Convention Record*, Part II, Vol. 11 (1963), pp. 83-90.

4. G. Biorci, P. P. Civalleri, G. Fiorio, "Degeneracy in n-Port Networks" *Proceedings, First Annual Allerton Conference on Circuit and System Theory* (University of Illinois) 1963, pp. 537-561.

5. R. M. Foster, "Topologic and Algebraic Considerations in Network Synthesis," *Proceedings, Symposium on Modern Network Synthesis* (New York) Vol. 1 (1952), pp. 8-18.

6. M. B. Reed, S. Seshn, "On Cut-Sets of Electrical Networks," *Proceedings, Second Midwest Symposium on Circuit Theory* (Michigan) 1956, pp. 1-13.

7. I. Cederbaum, "Invariance and Mutual Relations of Electrical Network Determinants," *Journal of Mathematics and Physics*, Vol. 23 (April 1954), pp. 236-244.

8. V. Baranov, "Méthode de Calcul des Réseaux d'Impédances," *Revue Générale d'Electricité*, Vol. 37 (March, 1935), pp. 339-351.

9. P. P. Civalleri, *On Generalized Cohn's Theorem*, to be published.

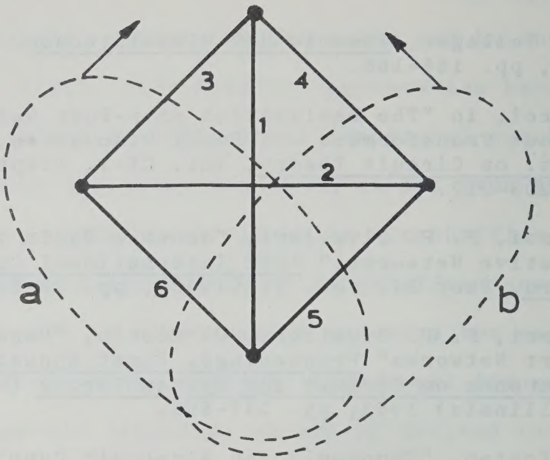


Fig. 1

Example of non fundamental cut-sets

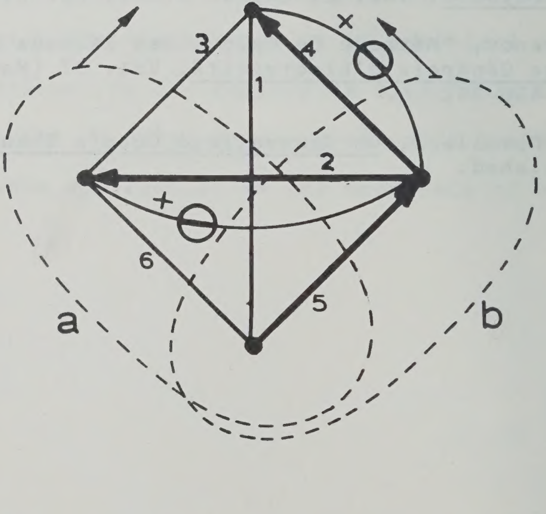


Fig. 2

Experimental arrangement to measure the mutual admittance between two non fundamental cut-sets

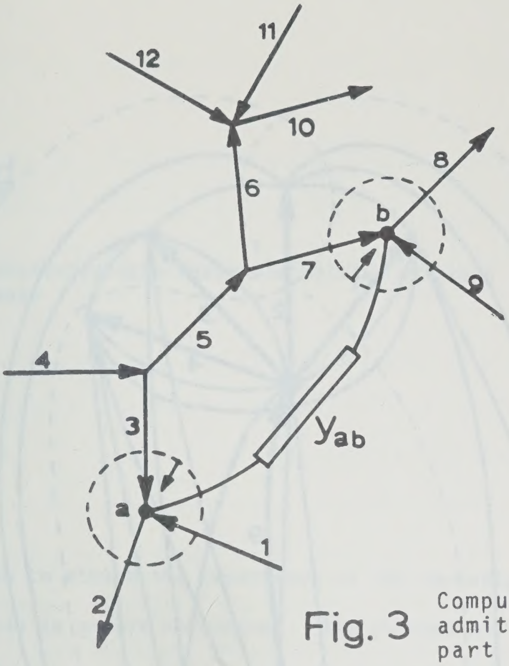

11
12
10
8
6
b
7
9
5
4
3
y_{ab}
a
1
2

Fig. 3 Computation of a branch admittance in terms of part admittances

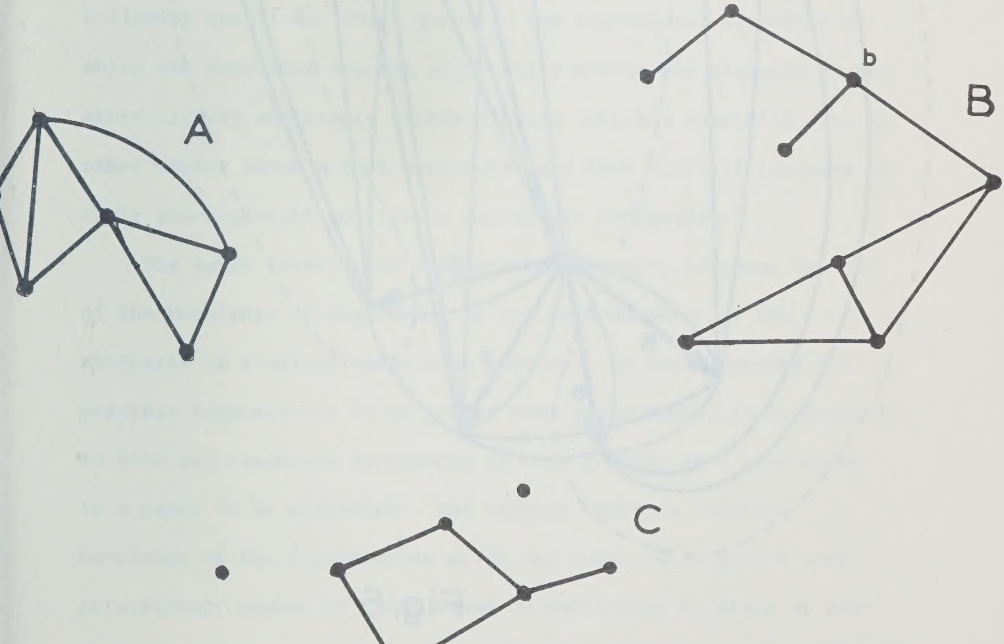


Fig. 4

Intuitive representation of classes A B C

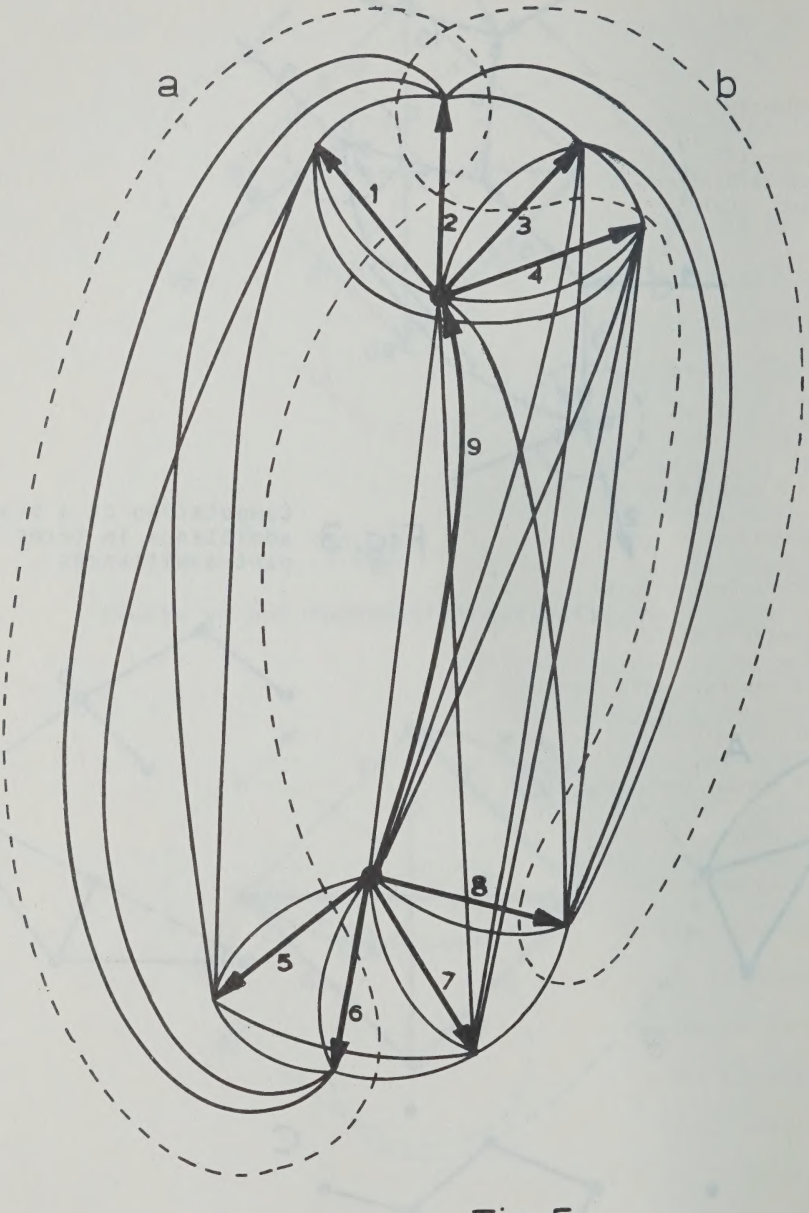

a
b
c
1
2
3
4
5
6
7
8
9

Fig. 5

Example of linear degeneracy

DISCUSSION

G. Biorci
Istituto Elettrotecnico Nazionale Galileo Ferraris
Torino, Italy

I wish to stress the importance of the investigation on degeneracies in _n_-port structures. The problem has a considerable interest "per se", since its exhaustive solution answers the following questions: which _n_-ports, the topological structure of which are such that one can arbitrarily assign all elements of the short-circuit admittance matrix (inside suitable domains)? Or, in other words: which _n_-port structures are such that all elements of their short-circuit admittance matrix are independent?

The major interest of the problem, however, concerns the use of the knowledge of degeneracy in the hard question of the synthesis of single-element kind _n_-ports. In fact, knowing all possible degeneracies in an _n_-port with (_n_+2) nodes, it is possible to find all canonical structures of such _n_-port, as I have shown in a paper to be published. And without doubt, a complete knowledge of the degeneracies which may occur in an _n_-port with an arbitrary number of nodes would be sufficient to prove or disprove my conjecture on the minimum number of branches. Therefore

I propose the following works, some of which may be good subjects for theses:

1.- About linear degeneracies: find how many branches an n-port based on (n+p) nodes (p larger than 2) must have in order to be unaffected by linear degeneracies. Find also how these branches must be located.

2.- About linear degeneracies: making use of Civalleri's theorems on linear degeneracies, find a systematic procedure to identify all linear degeneracies in a given n-port.

3.- About non-linear degeneracies: investigate various classes of non-linear degeneracies; find which of them do not imply linear degeneracies; find whether the set of known classes is complete or not (that is, whether an n-port which is not linearly degenerate and does not belong to any of the known classes of non-linear degeneracies, is non-degenerate or not).

4.- About the conjecture: find if there is at least one n-port with more than (n+2) nodes which is non-degenerate and has the minimum number of branches.

COMPUTERS AS AN AID
IN THE DEVELOPMENT OF TRANSMISSION CIRCUITS

C. W. Rosenthal and H. Simon
Bell Telephone Laboratories, Inc.
Murray Hill, N.J. and North Andover, Mass.

INTRODUCTION

Digital computers are used extensively in the development and design of transmission circuits. By transmission circuits we mean the building blocks of a transmission system, that is "black boxes" containing linear and nonlinear electrical elements and having inputs and outputs. Typical examples of transmission circuits are filters, modulators, regenerators and amplifiers. Our experience with the development of computer programs for the analysis, synthesis and optimization of transmission circuits started in the early 1950's. Today, approximately 150 computer programs are available and in constant use. The programs are an integral part of our system development capability and are representative, we believe, of the computational tools required to rapidly, efficiently, and adequately design systems. The focus will be on some important computer applications for frequency and time domain analysis of linear active and passive networks, synthesis of filters and optimization of circuits.

ANALYSIS OF TRANSMISSION CIRCUITS

Computer programs for analysis of networks evaluate the performance of a circuit, given its configuration and element values. With the aid of such programs, designs can be verified rapidly before laboratory experiments are made. In this section, we will give examples for the use of three analysis programs. Two of these, the Transmission Network Analysis Program and the Nodal Analysis Program, perform a frequency domain analysis of linear, active, and passive transmission circuits. The third program performs time domain analysis of linear networks.

Frequency Domain Analysis

Both the Transmission Network Analysis and Nodal Analysis Programs analyze the response of a circuit at specified frequencies. The Transmission Network Analysis Program is restricted to ladder networks whereas the Nodal Analysis Program can handle arbitrary configurations.

As a typical example of the use of these programs, we present the evaluation of a temperature regulating amplifier (Fig. 1). The impedance Z in the feedback path is the input impedance of the regulating network shown in Fig. 2. This regulating network is terminated in a thermistor which will change the input impedance and thus the feedback of the amplifier as a function of temperature. The objective of the computer study was to determine the gain characteristic of the amplifer at various temperatures of the thermistor. The amplifier is described to the Nodal Analysis Program by numbering the nodes as shown in Fig. 1 and then specifying the types and values of the elements between the nodes as in Fig. 3.

The Nodal Analysis Program can handle circuits with up to 24 nodes. Consequently, the input impedance of the regulating network for a specific value of the thermistor resistance, R_{TH}, was calculated using the Transmission Network Analysis Program. Fig. 4 shows the result of the impedance calculation. The values of the complex impedance Z representing the regulating network were read into the Nodal Analysis Program (see Code READ Z in Fig. 3).

The results of the gain calculation of the regulating amplifier for three different values of R_{TH} are shown in Fig. 5. The programs are flexible enough to permit ready changes in the circuit configuration. A thorough computer study of the circuit performance can be made in approximately one minute at a cost of about $12 per run on an IBM 7094.

Time Domain Analysis of Transmission Circuits

With the advent of digital transmission systems, flexible computer programs for transient analysis of transmission circuits have become increasingly important. In this section, we give an example of the application of a program capable of performing time domain analysis of linear networks of arbitrary configurations.

The problem at hand is to find the step response of the low-pass filter shown in Fig. 6. The method used in this computer program is based upon the numerical solution of linear first order differential equations, the state space equations. The current and voltage variables relating input and output have to be chosen so that the network can be described by first order linear differential equations. Aside from this restriction, the program can handle any linear network.

The system of differential equations describing the low-pass filter of Fig. 6 is given in matrix form in Fig. 7. The input data, as read into the computer, are shown in Fig. 8. Starting in the fourth line, the entries of the matrix shown in Fig. 7 are listed. The numbers in the first and second columns of Fig. 8 designate row and column of the entry. All entries not listed are assumed to be zero. The computer calculates the response of the output of the low-pass filter to a step at the input at discrete values of time. A microfilm plot of the output as a function of time is shown in Fig. 9.

SYNTHESIS OF FILTERS

Computer programs for synthesizing circuits obtain the configuration and element values of a circuit which will meet the designer's specifications. Programs, based on the insertion loss method,(1) now exist for the design of lumped parameter, linear passive circuits. The networks which can be synthesized use only reactive elements. Resistors occur only at the source and the load. An illustrative low-pass filter circuit, shown in Fig. 10, consists of inductors and capacitors in a ladder arrangement, with resistors for the load and the source.

Heretofore, computers have been used for piecemeal design of filters. Successive design stages have been programmed and executed separately. The program considered here is an integrated sequence of many separate programs to automatically design low-pass filters. Its flow is shown in Fig. 11. Specifications, as given in Fig. 12, are accepted and operated on to produce transmission characteristics. These are used to obtain the input impedance at the load and source ends. Finally, the configuration and the element values are produced. A predistortion program alters the network to account for predictable losses in the reactive components. Analysis programs operate upon the design data to produce tables and graphs describing the response of the network. Formats of input and output were selected to accommodate the engineer-user (Figs. 13 through 16).

Computation time depends upon the complexity of the network. The network in the above mentioned illustrations was designed and analyzed in approximately 18 seconds on an IBM 7094 at a cost of approximately $3.25. The total elapsed time between the supply of the punched cards to the computation center and the return of the data to the engineer's office would be two hours. This program provides rapid turn-around time, low cost, precise and accurate designs. These features permit our engineers to obtain more complex designs with far less effort.

Figures 17 and 18 show the computer design of a microwave filter. Starting from the requirements, the computer calculates the dimensions of the waveguide filter (Fig. 17). In addition, the drawing shown in Fig. 18 is prepared by the microfilm plotter. The design was completed in about 5 seconds on an IBM 7094 for $0.90.

OPTIMIZATION OF CIRCUITS

Optimization programs are useful in two respects. First, to improve upon the trial designs that engineers obtain by experience and intuition and second, to produce designs while considering practical constraints like limited power dissipation, bounds on the element values, and limits on the number of components. These constraints are not adequately considered in present synthesis procedures.

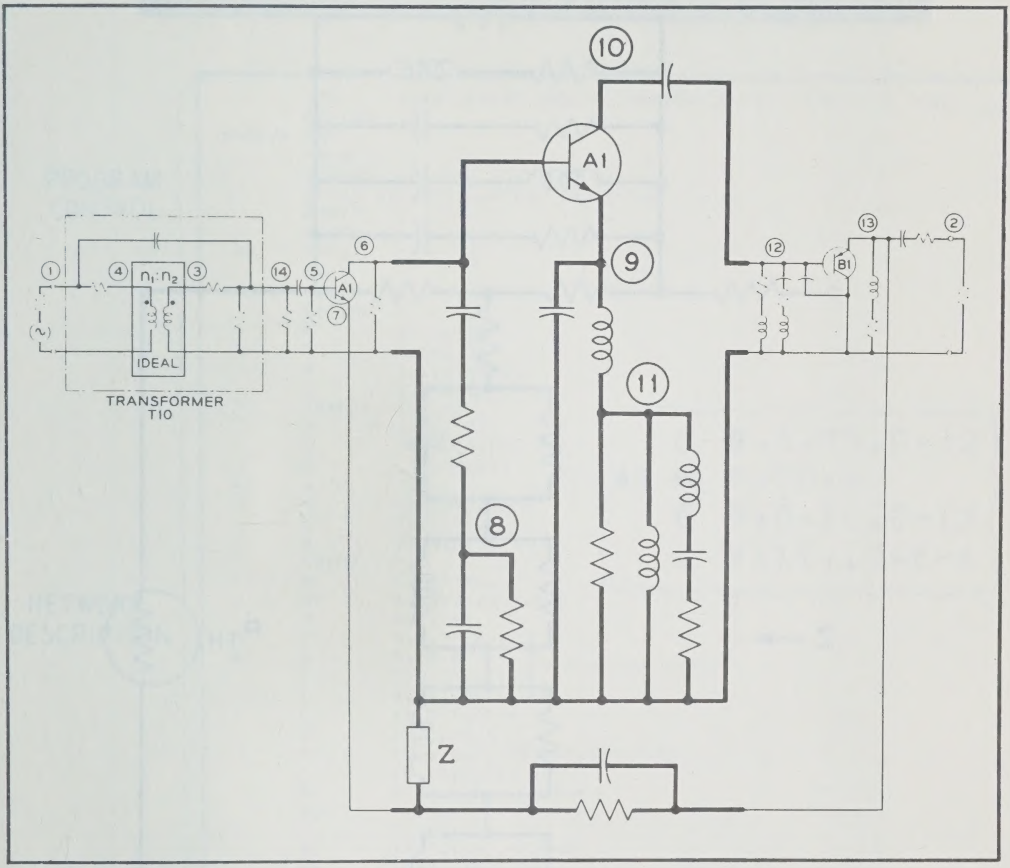
The performance required of a network and the performance obtained from the trial network are shown in Fig. 19. These differ at each of the specified frequencies f_1 through f_6. A systematic procedure is needed for reducing these differences. The sum of the squares of the differences is minimized primarily because it leads to the most tractable optimization procedures. One method used is called the steepest descent method or the "lost in the woods" method.(2) A person who is lost in the woods is advised to proceed downhill at all times because civilization generally locates in the low parts of a terrain. The optimization procedure forces us to proceed in a downhill fashion on the sum of the squares of the differences. Figure 20 is a crude pictorial indication of what is occurring.

The minimum sum of squared differences may not be obtained because of the practical constraints or because the minimum function contains a hollow. In this case, one must climb up before going to a lower point. Ascent cannot at present be handled by our systematic procedures. Figure 21 shows the program flow for the optimization of the performance of a network. Figures 22 and 23 illustrate the way in which a delay equalizing network has been improved through the use of this procedure. The computer work was completed in about 20 seconds on an IBM 7094 for about $3.60. Many network types may be optimized with our present programs and many aspects of these networks may be considered for optimization.

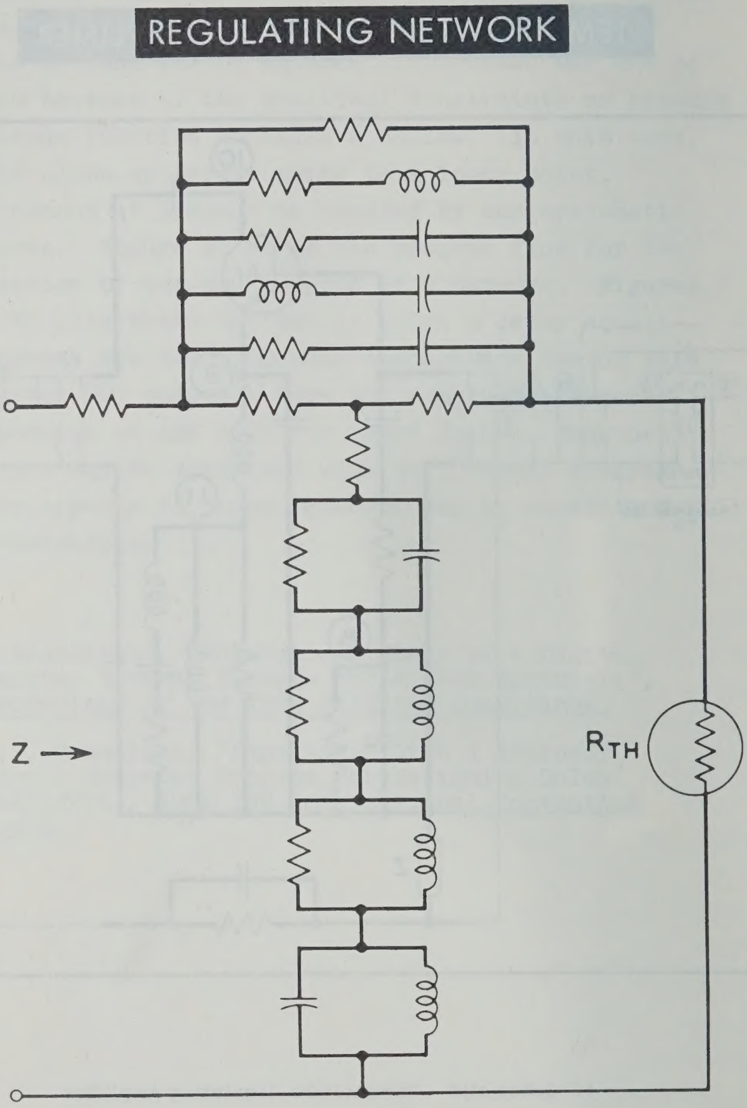
(1) G. Szentirmai, "Theoretical Basis of a Digital Computer Program Package for Filter Synthesis", Proceedings of the 1963 Allerton Conference.

(2) C. L. Semmelman, "Experience with a Steepest Descent Computer Program for Designing Delay Equalizers", 1962 IRE International Convention Record.

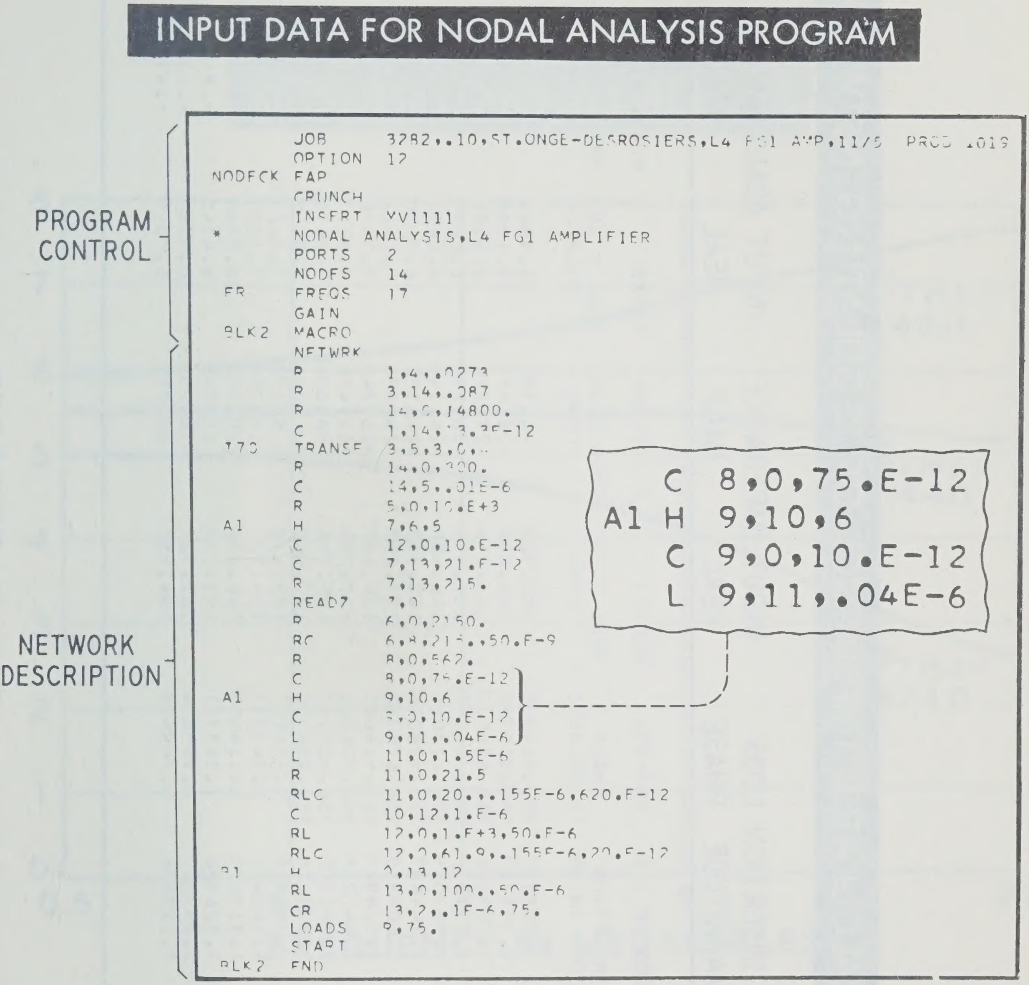
TEMPERATURE REGULATING AMPLIFIER


10
A1
9
11
8
Z
1
4
$n_1 : n_2$
3
14
5
A1
6
7
(~)
IDEAL
TRANSFORMER
T10
12
B1
13
2

1. Schematic - Temperature Regulating Amplifier


REGULATING NETWORK
Z →
R_TH

2. Schematic - Regulating Network


INPUT DATA FOR NODAL ANALYSIS PROGRAM
PROGRAM CONTROL
NETWORK DESCRIPTION

```
        JOB      3282,.10,ST.ONGE-DESROSIERS,L4 FG1 AMP,11/5  PROD .019
        OPTION   12
NODECK  FAP
        CRUNCH
        INSERT   YV1111
*       NODAL ANALYSIS,L4 FG1 AMPLIFIER
        PORTS    2
        NODES    14
 FR     FREQS    17
        GAIN
 BLK2   MACRO
        NETWRK
        R        1,4,.0273
        R        3,14,.087
        R        14,0,14800.
        C        1,14,13.3E-12
 T70    TRANSF   3,5,3,0,-
        R        14,0,300.
        C        14,5,.01E-6
        R        5,0,10.E+3
 A1     H        7,6,5
        C        12,0,10.E-12
        C        7,13,21.E-12
        R        7,13,215.
        READ7    7,0
        R        6,0,2150.
        RC       6,8,215.,50.E-9
        R        8,0,562.
        C        8,0,75.E-12
 A1     H        9,10,6
        C        9,0,10.E-12
        L        9,11,.04E-6
        L        11,0,1.5E-6
        R        11,0,21.5
        RLC      11,0,20.,.155E-6,620.E-12
        C        10,12,1.E-6
        RL       12,0,1.E+3,50.E-6
        RLC      12,0,61.9,.155E-6,20.E-12
 B1     H        9,13,12
        RL       13,0,100.,50.E-6
        CR       13,2,.1E-6,75.
        LOADS    9,75.
        START
 BLK2   END
```

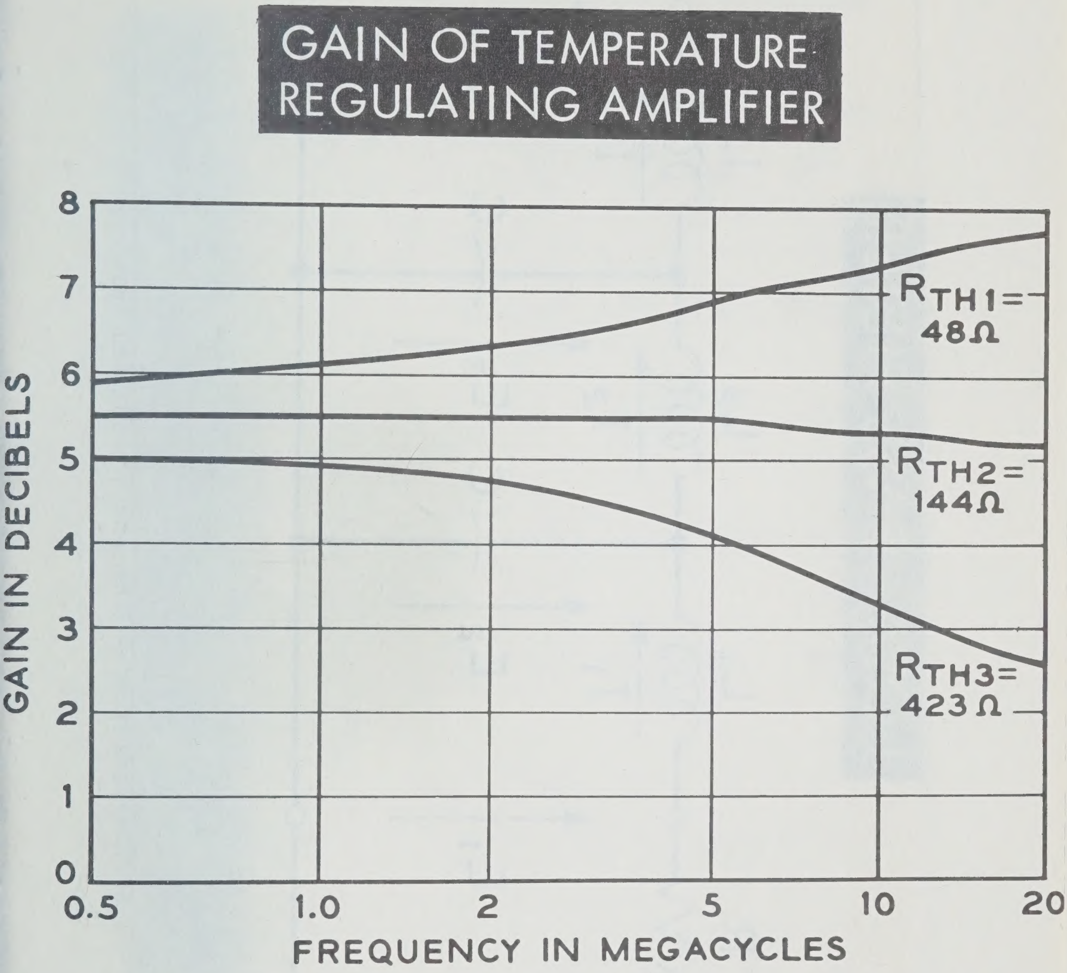
```
   C 8,0,75.E-12
A1 H 9,10,6
   C 9,0,10.E-12
   L 9,11,.04E-6
```

3. Input Data for Nodal Analysis Program

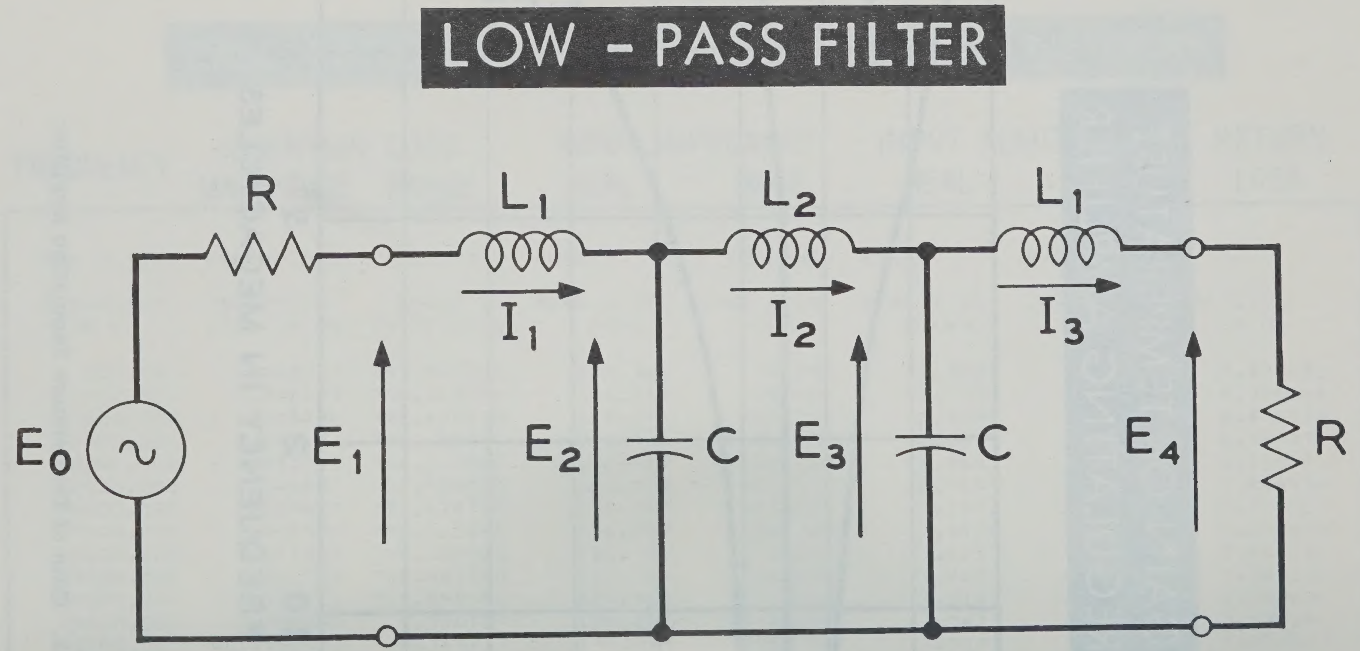
CALCULATED IMPEDANCE OF REGULATING NETWORK

FREQUENCY	INSERTION LOSS MAGNITUDE	INSERTION LOSS PHASE	INPUT IMPEDANCE REAL	INPUT IMPEDANCE IMAG.	INPUT ADMITTANCE REAL	INPUT ADMITTANCE IMAG.	RETURN LOSS

```
NOV 7,1963   CODERD S O   ROB=872        RUN=23       F COLLINS   R1=80       R=47.96

FREQUENCY    INSERTION      PHASE        REAL Z IN    IMAG Z IN   REAL Y IN   IMAG Y IN   RET.LOSS
  IN KC      LOSS IN DB   IN DEGREES       OHMS         OHMS       MILMHOS     MILMHOS     IN DB

   500.000    7.388157    353.829800     202.361       -4.795      4.9389       .1170    8.831430
  1000.000    5.395183    344.260740     194.254      -18.654      5.1009       .4898    8.396666
  5000.000    2.709196    348.676630     163.254      -27.294      5.9589       .9962    6.876978
  7000.000    2.292537    350.266050     155.054      -26.693      6.2637      1.0783    6.499375
 10000.000    1.933059    352.213660     146.486      -24.140      6.6461      1.0953    6.114479
 12500.000    1.744687    353.526010     141.333      -21.532      6.9150      1.0535    5.886936
 15000.000    1.609919    354.678710     137.297      -18.676      7.1512       .9727    5.710567
 17000.000    1.528929    355.526680     134.697      -16.245      7.3176       .8825    5.597806
 20000.000    1.441268    356.706020     131.704      -12.436      7.5257       .7106    5.468731
 25000.000    1.361708    358.465190     128.794       -6.008      7.7475       .3614    5.344035
 30000.000    1.339815    359.998550     127.960         .013      7.8149      -.0008    5.308446
 35000.000    1.356846      1.334107     128.612        5.270      7.7623      -.3181    5.336285
 50000.000    1.542433      4.343217     134.923       15.739      7.3121      -.8530    5.608417
 65000.000    1.813633      6.124553     142.426       20.002      6.8854      -.9670    5.938523
 95000.000    2.326366      7.343701     153.206       20.345      6.4140      -.8517    6.427561
145000.000    2.853372      6.785814     161.327       16.254      6.1363      -.6182    6.809831
195000.000    3.113426      5.749418     164.555       12.891      6.0399      -.4732    6.965621
```



5. Gain of Temperature Regulating Amplifier


LOW - PASS FILTER
R
L1
L2
L1
I1
I2
I3
E0
E1
E2
C
E3
C
E4
R

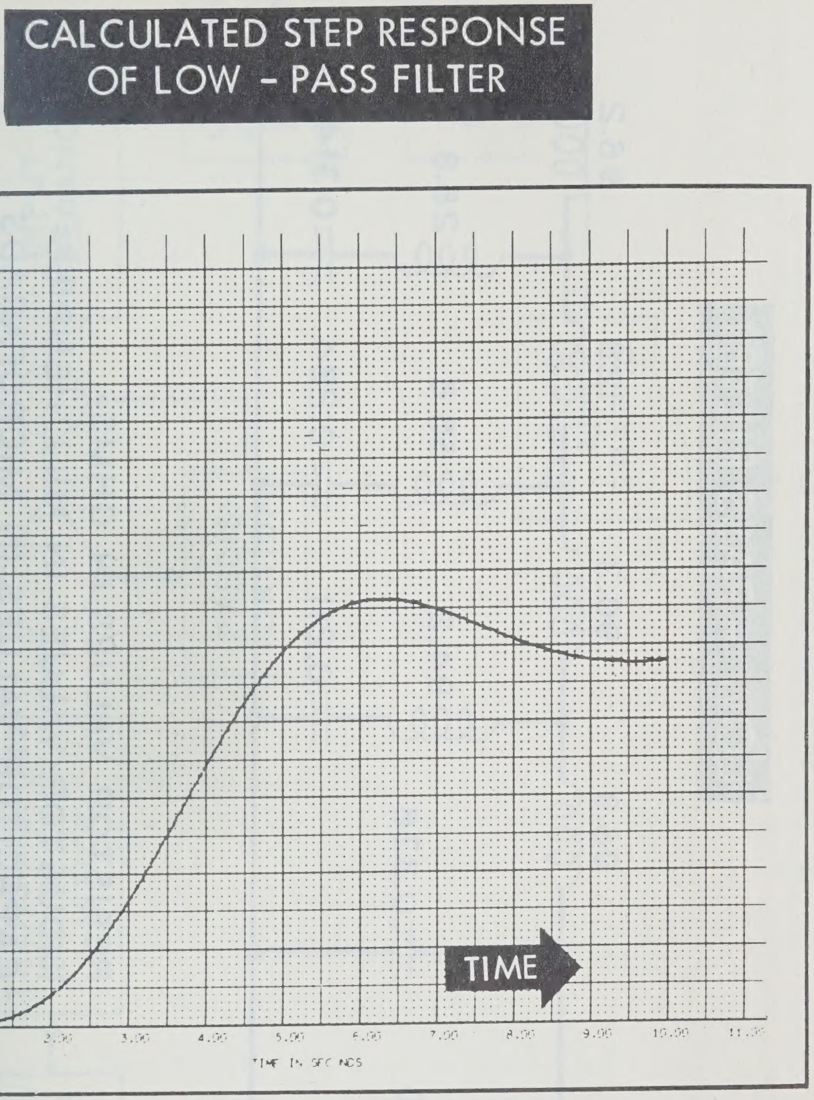
6. Low-pass Filter -- Schematic (Time Domain Analysis)

MATRIX FORM OF DIFFERENTIAL EQUATIONS DESCRIBING THE FILTER

$$\frac{d}{dt}\begin{bmatrix} I_1 \\ E_2 \\ I_2 \\ E_3 \\ I_3 \end{bmatrix} = \begin{bmatrix} -\frac{R}{L_1} & -\frac{1}{L_1} & 0 & 0 & 0 \\ \frac{1}{C} & 0 & -\frac{1}{C} & 0 & 0 \\ 0 & \frac{1}{L_2} & 0 & -\frac{1}{L_2} & 0 \\ 0 & 0 & \frac{1}{C} & 0 & -\frac{1}{C} \\ 0 & 0 & 0 & \frac{1}{L_1} & -\frac{R}{L_1} \end{bmatrix} \begin{bmatrix} I_1 \\ E_2 \\ I_2 \\ E_3 \\ I_3 \end{bmatrix} + \begin{bmatrix} \frac{E_0}{L_1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

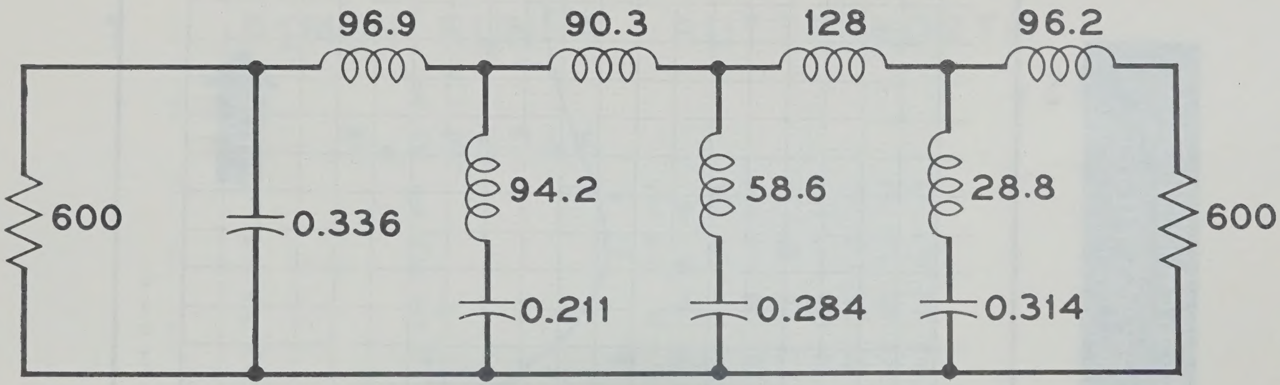
INPUT DATA FOR TRANSIENT ANALYSIS

```
1    SIMON RUN 3  BUTTERWORTH
     5    10                     .20
        3.236246
     1     1       -1.6181229
     1     2       -1.6181229
     2     1         .61804697
     2     3        -.61804697
     3     2                .5
     3     4               -.5
     4     3         .61804697
     5     4        1.6181229
     5     5       -1.6181229
                     9.9999999E+00
```



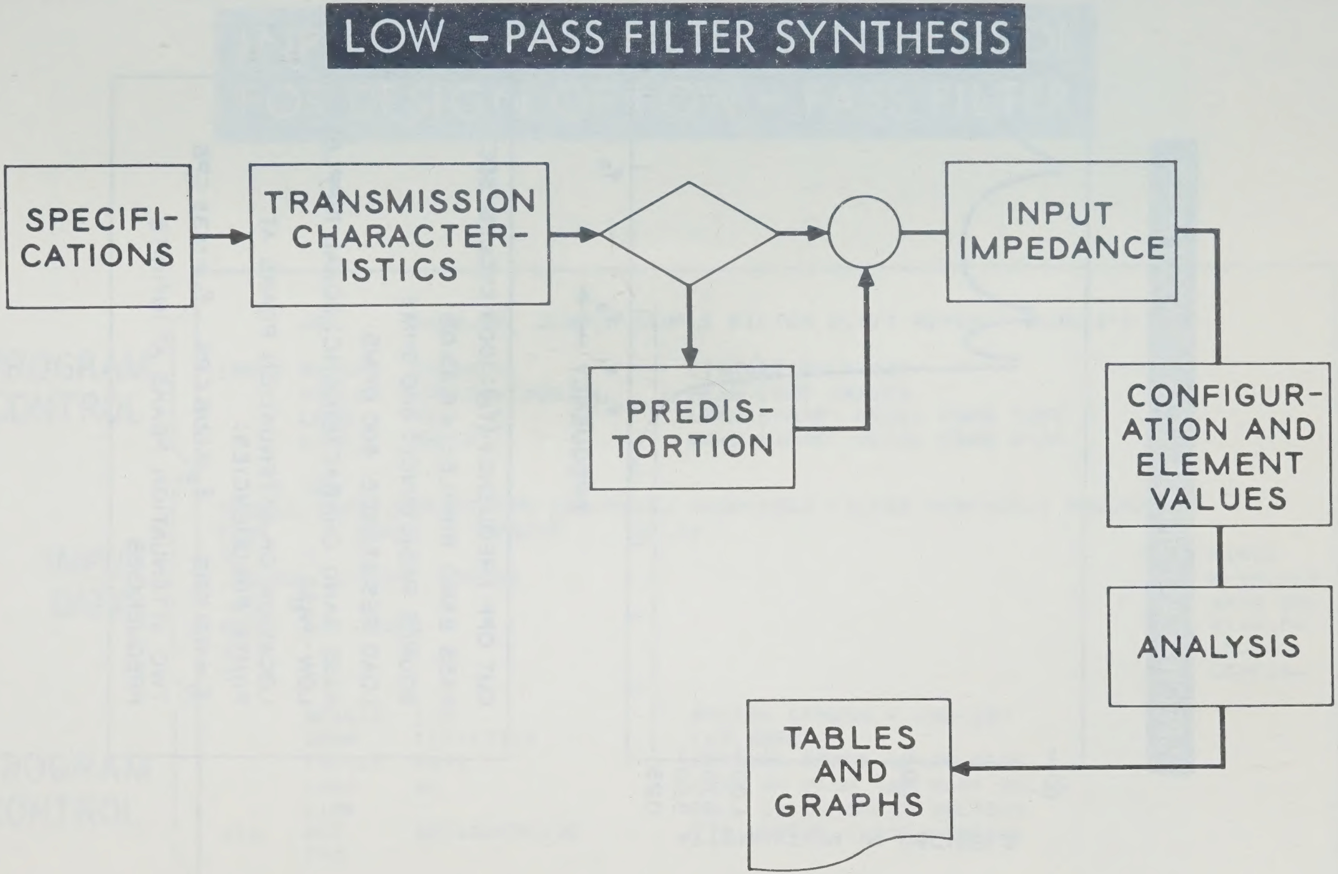
9. Calculated Step Response of Low-pass Filter

LOW - PASS FILTER


96.9
90.3
128
96.2
600
0.336
94.2
0.211
58.6
0.284
28.8
0.314
600

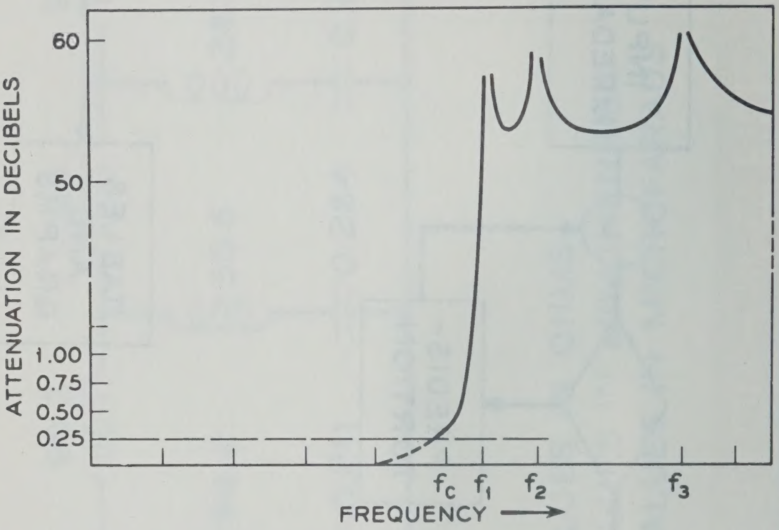
RESISTOR VALUES IN OHMS
INDUCTOR VALUES IN MILLIHENRIES
CAPACITOR VALUES IN MICROFARADS

10. Low-pass Filter (Synthesis)


LOW - PASS FILTER SYNTHESIS
SPECIFICATIONS
TRANSMISSION CHARACTERISTICS
PREDISTORTION
INPUT IMPEDANCE
CONFIGURATION AND ELEMENT VALUES
ANALYSIS
TABLES AND GRAPHS

11. Low-pass Filter Synthesis (Flow Chart)

SPECIFICATION OF LOW - PASS FILTER



CUT OFF FREQUENCY (f_c): 1000 CYCLES/SEC

PASS BAND RIPPLE: < 0.25 DB

SOURCE RESISTANCE: 600 OHMS

LOAD RESISTANCE: 600 OHMS

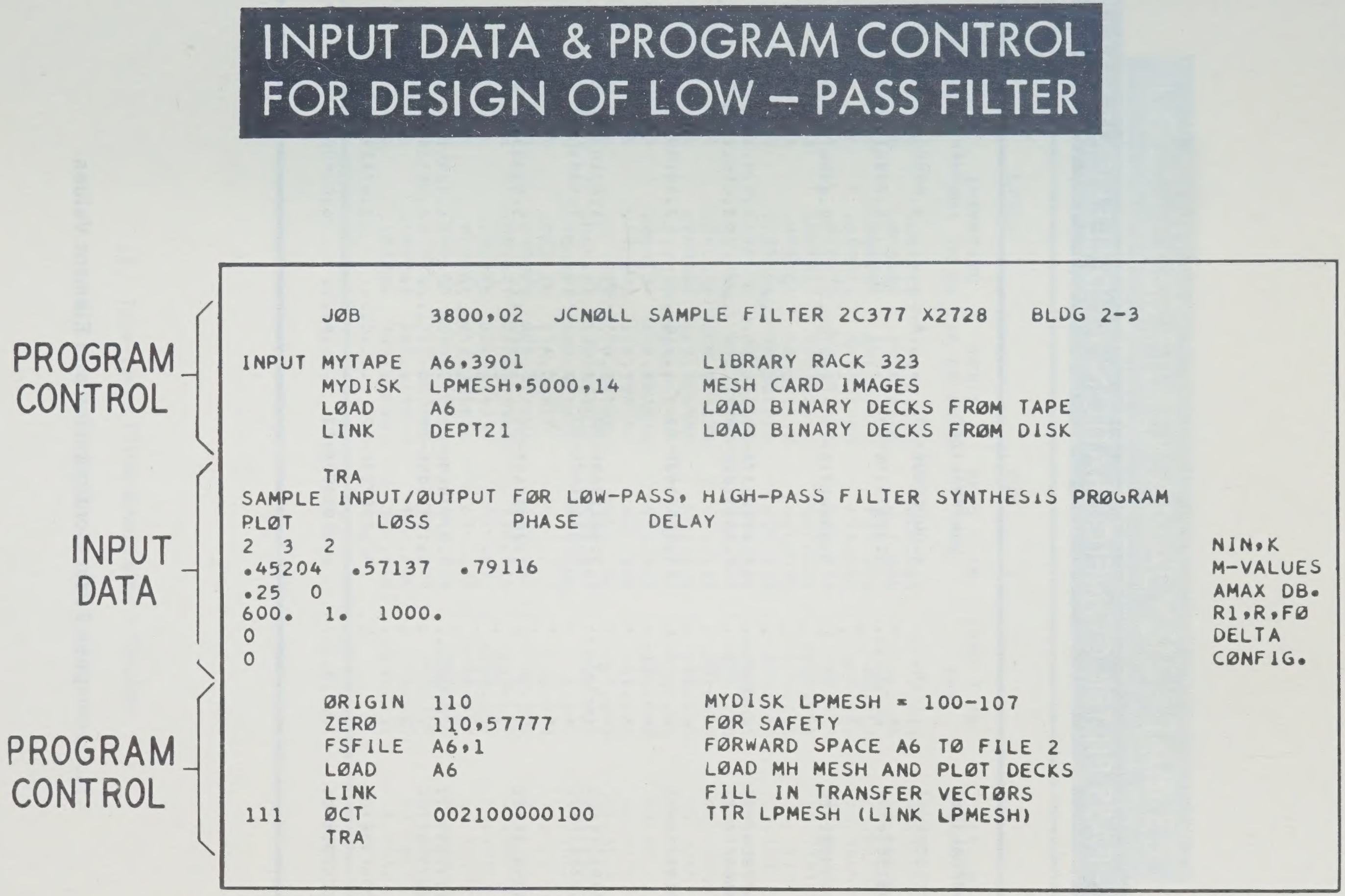
PASS BAND CHARACTERISTIC: EQUAL RIPPLE, LOW-PASS

LOCATION OF ATTENUATION PEAKS AT FINITE FREQUENCIES:

$f_1 = 1121$ CPS $f_2 = 1218$ CPS $f_3 = 1635$ CPS

TWO ATTENUATION PEAKS AT INFINITE FREQUENCIES

12. Specification of Low-pass Filter



INPUT DATA & PROGRAM CONTROL FOR DESIGN OF LOW – PASS FILTER

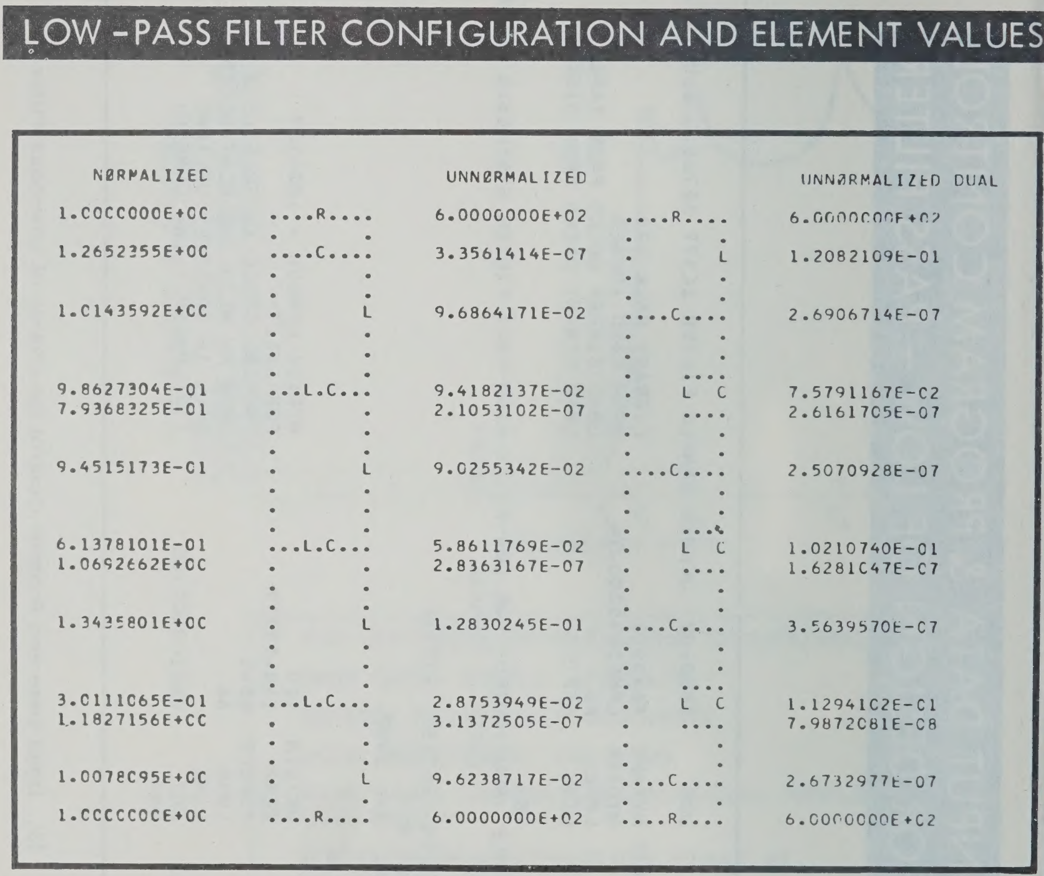
PROGRAM CONTROL

INPUT DATA

PROGRAM CONTROL

```
     JØB     3800,02  JCNØLL SAMPLE FILTER 2C377 X2728   BLDG 2-3

INPUT MYTAPE  A6,3901             LIBRARY RACK 323
     MYDISK  LPMESH,5000,14      MESH CARD IMAGES
     LØAD    A6                  LØAD BINARY DECKS FRØM TAPE
     LINK    DEPT21              LØAD BINARY DECKS FRØM DISK

     TRA
SAMPLE INPUT/ØUTPUT FØR LØW-PASS, HIGH-PASS FILTER SYNTHESIS PRØGRAM
PLØT      LØSS      PHASE     DELAY
2  3  2                                                          NIN,K
.45204  .57137  .79116                                           M-VALUES
.25  0                                                           AMAX DB.
600.  1.  1000.                                                  R1,R,FØ
0                                                                DELTA
0                                                                CØNFIG.

     ØRIGIN  110                 MYDISK LPMESH = 100-107
     ZERØ    110,57777           FØR SAFETY
     FSFILE  A6,1                FØRWARD SPACE A6 TØ FILE 2
     LØAD    A6                  LØAD MH MESH AND PLØT DECKS
     LINK                        FILL IN TRANSFER VECTØRS
111  ØCT     002100000100        TTR LPMESH (LINK LPMESH)
     TRA
```

13. Input Data and Program Control for Design of Low-pass Filter

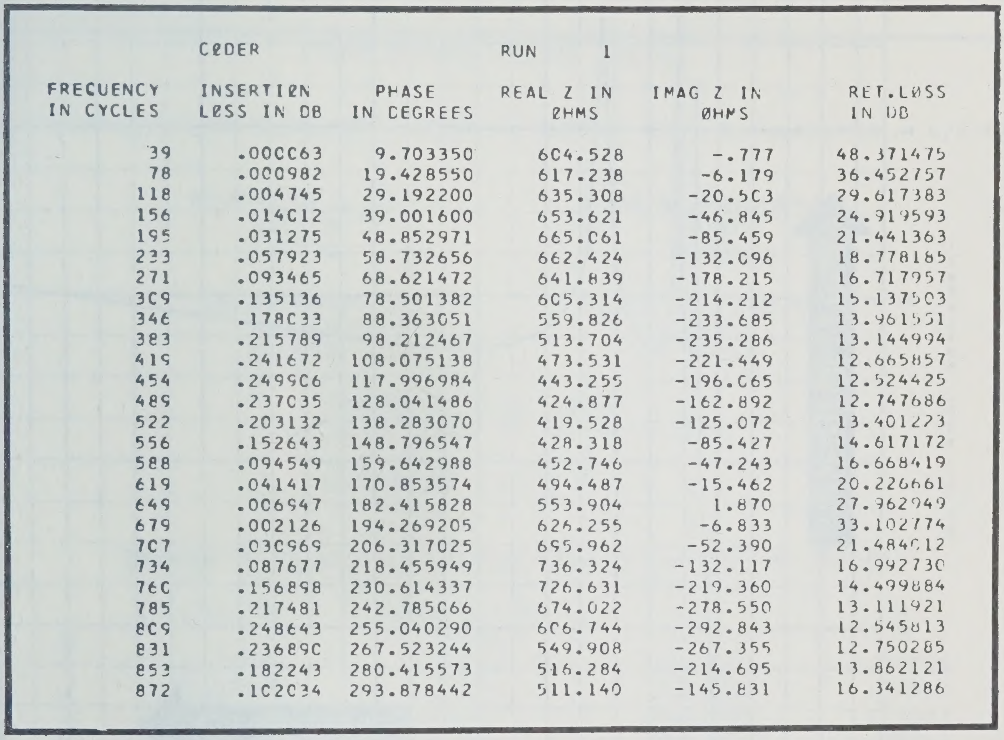


LOW -PASS FILTER CONFIGURATION AND ELEMENT VALUES

```
   NØRMALIZEC                     UNNØRMALIZED                  UNNØRMALIZED DUAL
1.COCC000E+0C     ....R....     6.0000000E+02    ....R....     6.0000C00F+02
                  .       .                      .       .
1.2652355E+0C     ....C....     3.3561414E-C7    .       L     1.2082109E-01
                  .       .                      .       .
                  .       .                      .       .
1.C143592E+CC     .       L     9.6864171E-02    ....C....     2.6906714E-07
                  .       .                      .       .
                  .       .                      .       .
                  .       .                      .    ....
9.8627304E-01     ...L.C...     9.4182137E-02    .    L  C     7.5791167E-C2
7.9368325E-01     .       .     2.1053102E-07    .    ....     2.6161705E-07
                  .       .                      .       .
                  .       .                      .       .
9.4515173E-C1     .       L     9.0255342E-02    ....C....     2.5070928E-07
                  .       .                      .       .
                  .       .                      .       .
                  .       .                      .    ....
6.1378101E-01     ...L.C...     5.8611769E-02    .    L  C     1.0210740E-01
1.0692662E+0C     .       .     2.8363167E-07    .    ....     1.6281C47E-C7
                  .       .                      .       .
                  .       .                      .       .
1.3435801E+0C     .       L     1.2830245E-01    ....C....     3.5639570E-C7
                  .       .                      .       .
                  .       .                      .       .
                  .       .                      .    ....
3.C111C65E-01     ...L.C...     2.8753949E-02    .    L  C     1.12941C2E-C1
1.1827156E+CC     .       .     3.1372505E-07    .    ....     7.9872C81E-C8
                  .       .                      .       .
                  .       .                      .       .
1.0078C95E+CC     .       L     9.6238717E-02    ....C....     2.6732977E-07
                  .       .                      .       .
1.CCCCCOCE+0C     ....R....     6.0000000E+02    ....R....     6.0000000E+C2
```

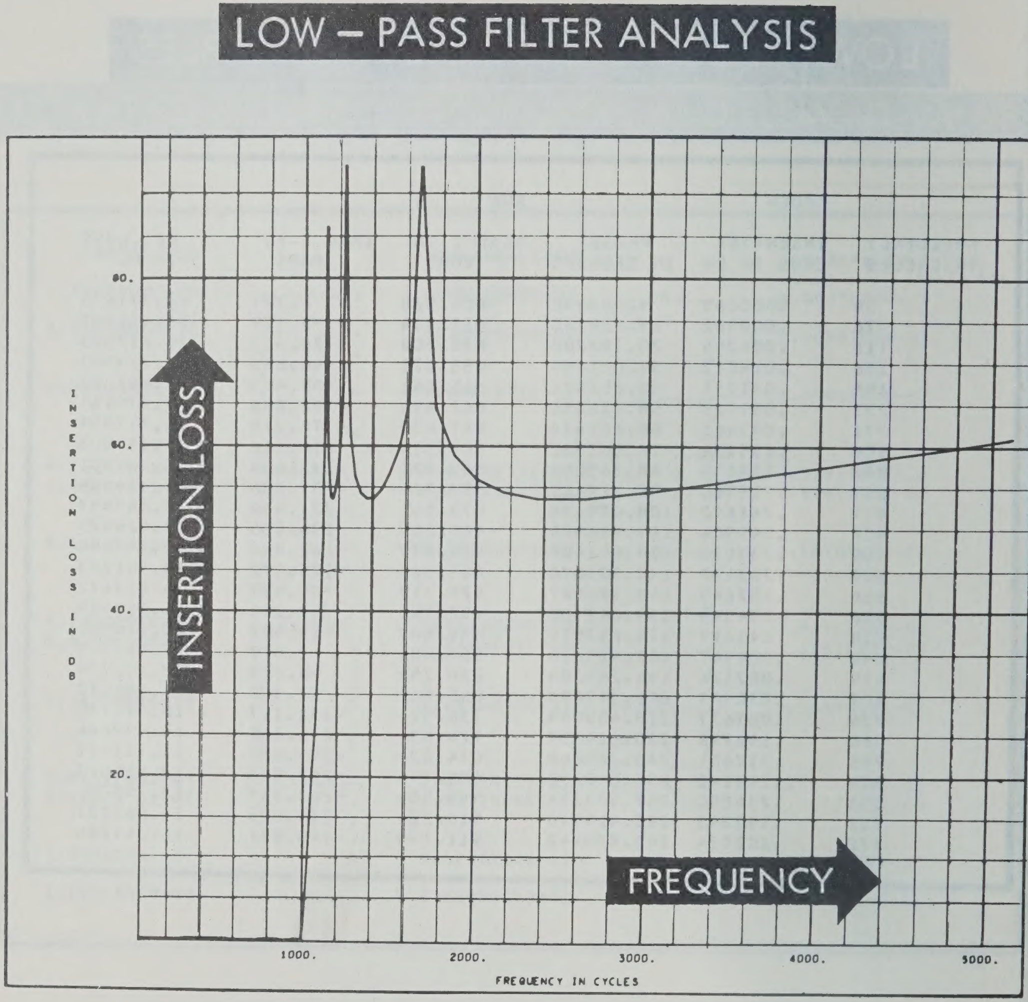
14. Low-pass Filter Configuration and Element Values

LOW - PASS FILTER ANALYSIS - TABLES

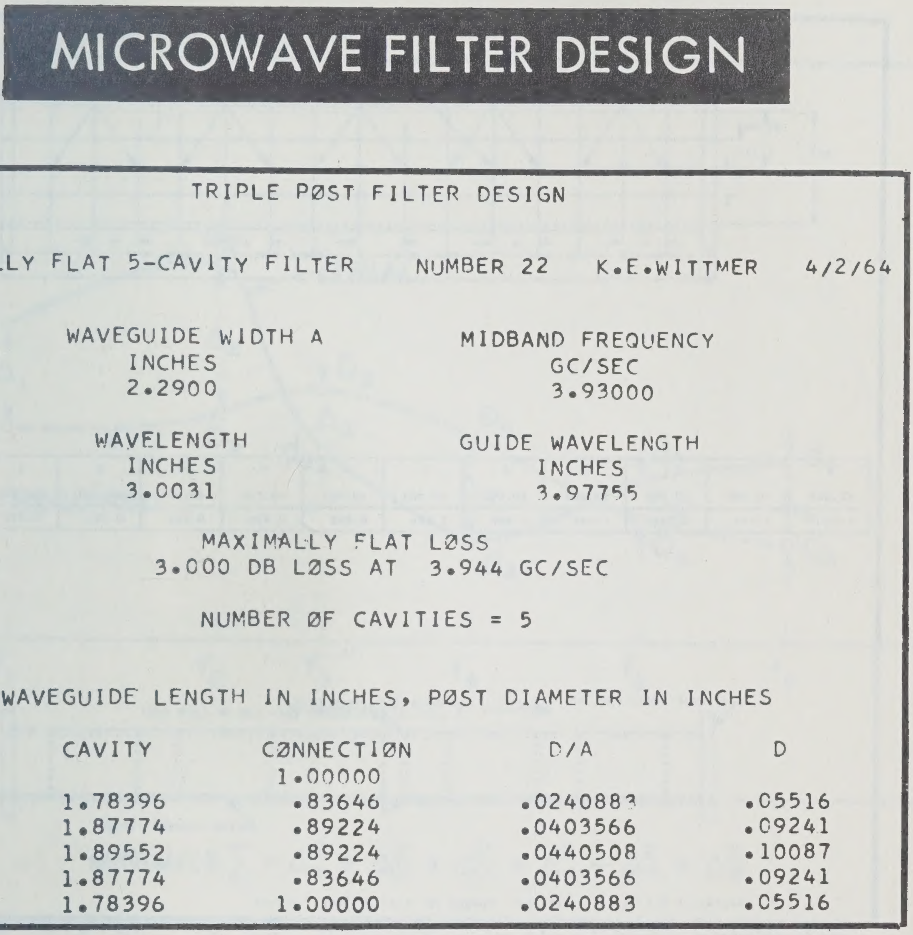


```
                CØDER                           RUN        1

FRECUENCY   INSERTIØN       PHASE       REAL Z IN    IMAG Z IN          RET.LØSS
IN CYCLES   LØSS IN DB   IN DEGREES       ØHMS          ØHMS             IN DB

       39     .000063      9.703350     604.528        -.777        48.371475
       78     .000982     19.428550     617.238       -6.179        36.452757
      118     .004745     29.192200     635.308      -20.503        29.617383
      156     .014012     39.001600     653.621      -46.845        24.919593
      195     .031275     48.852971     665.061      -85.459        21.441363
      233     .057923     58.732656     662.424     -132.096        18.778165
      271     .093465     68.621472     641.839     -178.215        16.717957
      309     .135136     78.501382     605.314     -214.212        15.137503
      346     .178033     88.363051     559.826     -233.685        13.961551
      383     .215789     98.212467     513.704     -235.286        13.144994
      419     .241672    108.075138     473.531     -221.449        12.665857
      454     .249906    117.996984     443.255     -196.065        12.524425
      489     .237035    128.041486     424.877     -162.892        12.747686
      522     .203132    138.283070     419.528     -125.072        13.401223
      556     .152643    148.796547     428.318      -85.427        14.617172
      588     .094549    159.642988     452.746      -47.243        16.668419
      619     .041417    170.853574     494.487      -15.462        20.226661
      649     .006947    182.415828     553.904        1.870        27.962949
      679     .002126    194.269205     626.255       -6.833        33.102774
      707     .030969    206.317025     695.962      -52.390        21.484012
      734     .087677    218.455949     736.324     -132.117        16.992730
      760     .156898    230.614337     726.631     -219.360        14.499884
      785     .217481    242.785066     674.022     -278.550        13.111921
      809     .248643    255.040290     606.744     -292.843        12.545813
      831     .236890    267.523244     549.908     -267.355        12.750285
      853     .182243    280.415573     516.284     -214.695        13.862121
      872     .102034    293.878442     511.140     -145.831        16.341286
```

15. Low-pass Filter Analysis -- Tables



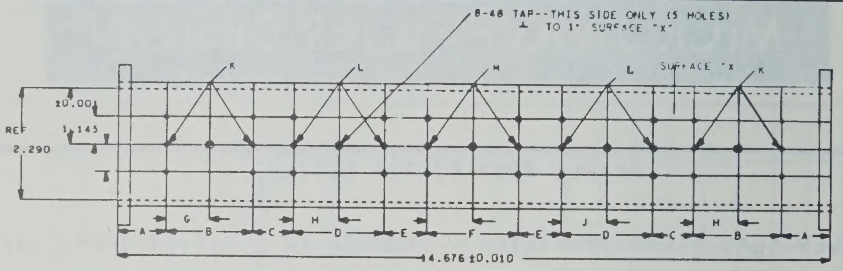
16. Low-pass Filter Analysis (Plot)



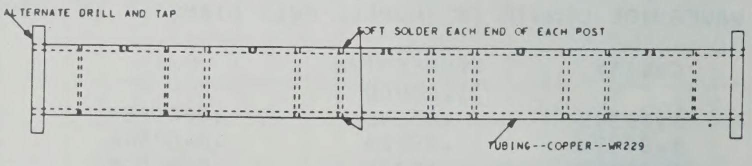
MICROWAVE FILTER DESIGN

```
                    TRIPLE PØST FILTER DESIGN

MAXIMALLY FLAT 5-CAVITY FILTER    NUMBER 22   K.E.WITTMER   4/2/64

          WAVEGUIDE WIDTH A              MIDBAND FREQUENCY
               INCHES                         GC/SEC
               2.2900                         3.93000

           WAVELENGTH                    GUIDE WAVELENGTH
               INCHES                         INCHES
               3.0031                         3.97755

                      MAXIMALLY FLAT LØSS
                 3.000 DB LØSS AT  3.944 GC/SEC

                      NUMBER ØF CAVITIES = 5


       WAVEGUIDE LENGTH IN INCHES, PØST DIAMETER IN INCHES

           CAVITY          CØNNECTIØN          D/A               D
 0                           1.00000
 1         1.78396            .83646         .0240883          .05516
 2         1.87774            .89224         .0403566          .09241
 3         1.89552            .89224         .0440508          .10087
 4         1.87774            .83646         .0403566          .09241
 5         1.78396           1.00000         .0240883          .05516
```

17. Microwave Filter Design


8-48 TAP--THIS SIDE ONLY (5 HOLES)
⊥ TO 1° SURFACE "X"
SURFACE "X"
K
L
M
L
K
±0.001
1.145
REF
2.290
G
H
J
H
A
B
C
D
E
F
E
D
C
B
A
14.676 ±0.010

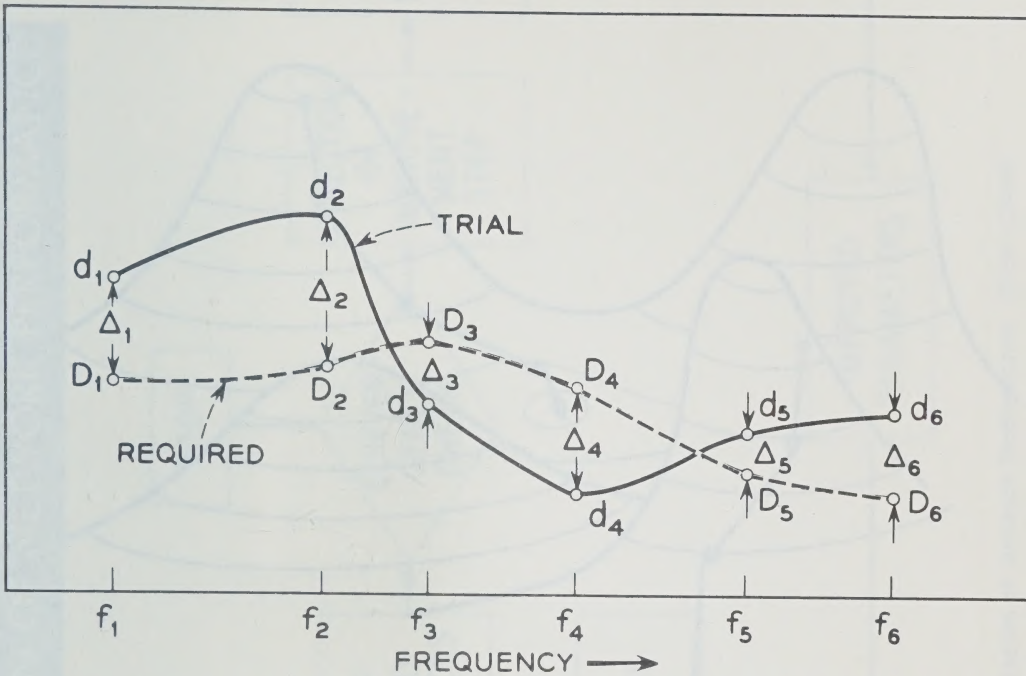
A	B	C	D	E	F	G	H	J	K	L	M
±0.005	±0.005	±0.002	±0.005	±0.002	±0.005	±0.005	±0.005	±0.005	±0.000	±0.000	±0.000
1.000	1.784	0.836	1.878	0.892	1.896	0.892	0.939	0.948	0.055	0.092	0.101


ALTERNATE DRILL AND TAP
SOFT SOLDER EACH END OF EACH POST
TUBING--COPPER--WR229

MAXIMALLY FLAT 5-CAVITY FILTER NUMBER 22 K.E. WITTMER 4/2/64

18. Drawing of Triple Post Filter

PERFORMANCE SPECIFICATIONS

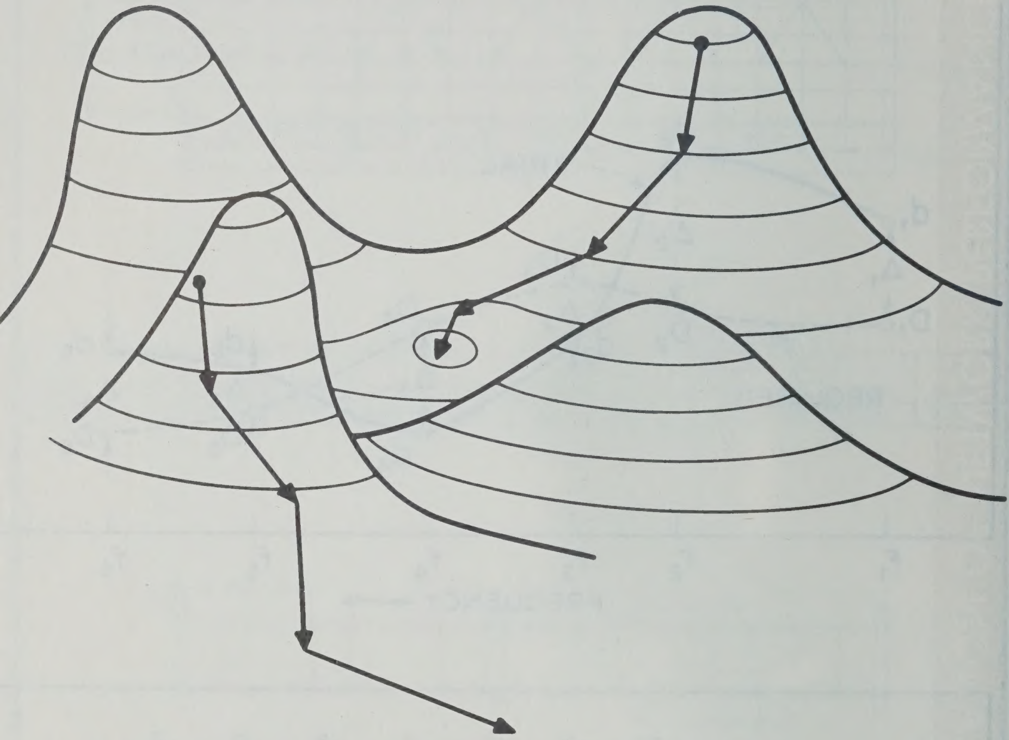

d₁
d₂
TRIAL
Δ₁
Δ₂
D₁
D₂
D₃
Δ₃
d₃
D₄
Δ₄
d₄
d₅
Δ₅
D₅
d₆
Δ₆
D₆
REQUIRED
f₁
f₂
f₃
f₄
f₅
f₆
FREQUENCY →

MINIMIZE $\sum = \Delta_1^2 + \Delta_2^2 + \Delta_3^2 + \Delta_4^2 + \Delta_5^2 + \Delta_6^2$

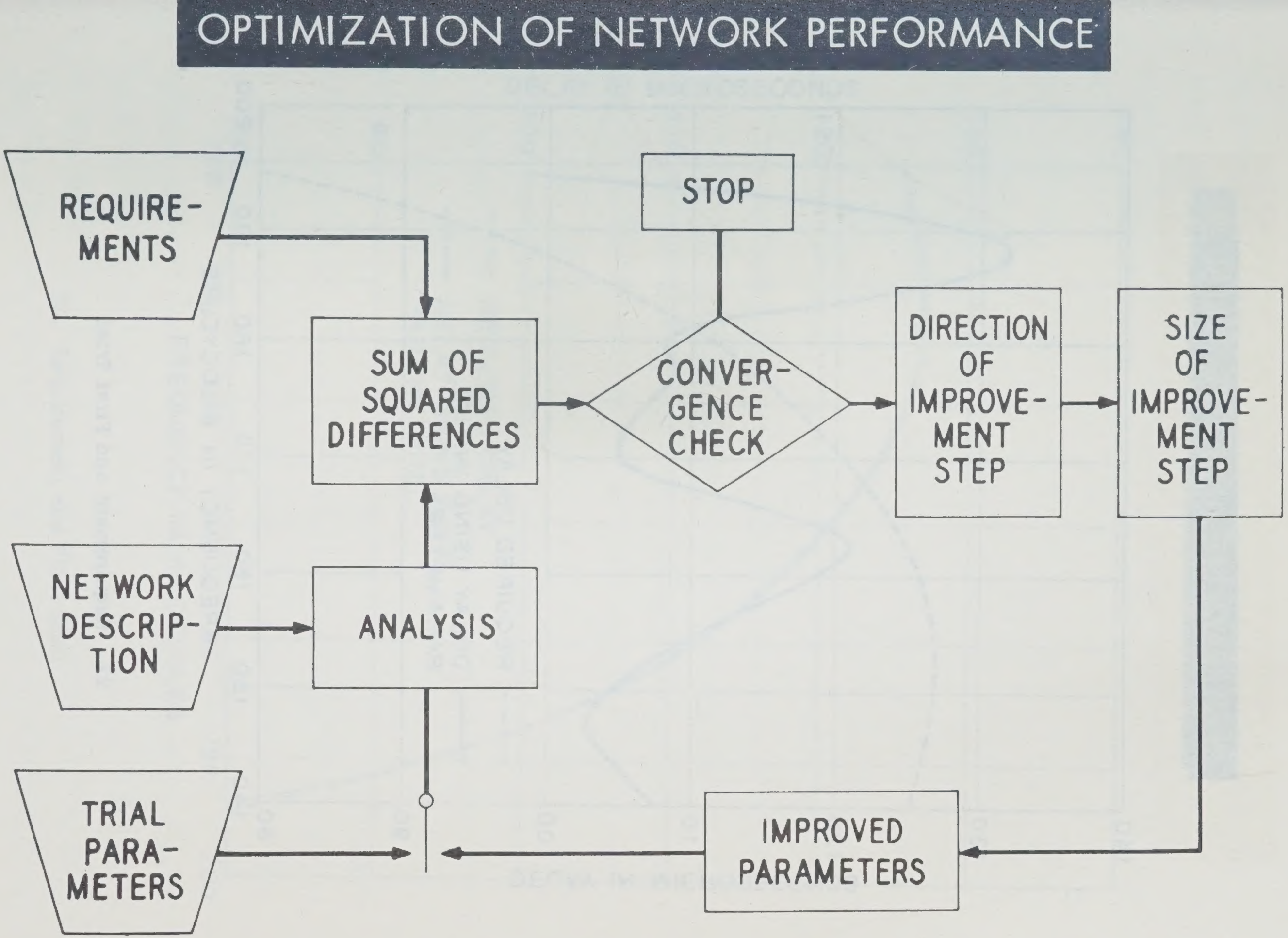
BY SYSTEMATICALLY VARYING THE NETWORK ELEMENTS

19. Performance Specifications (Steepest Descent)

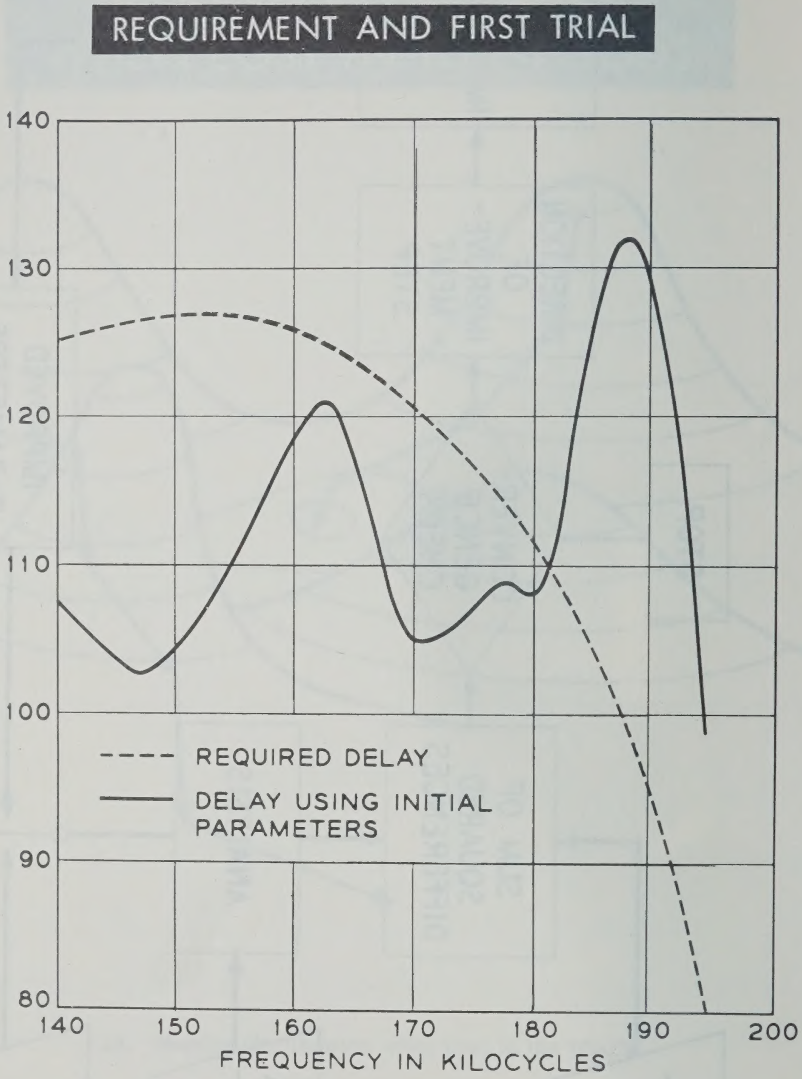
SEEKING CIVILIZATION WHEN LOST IN THE WOODS



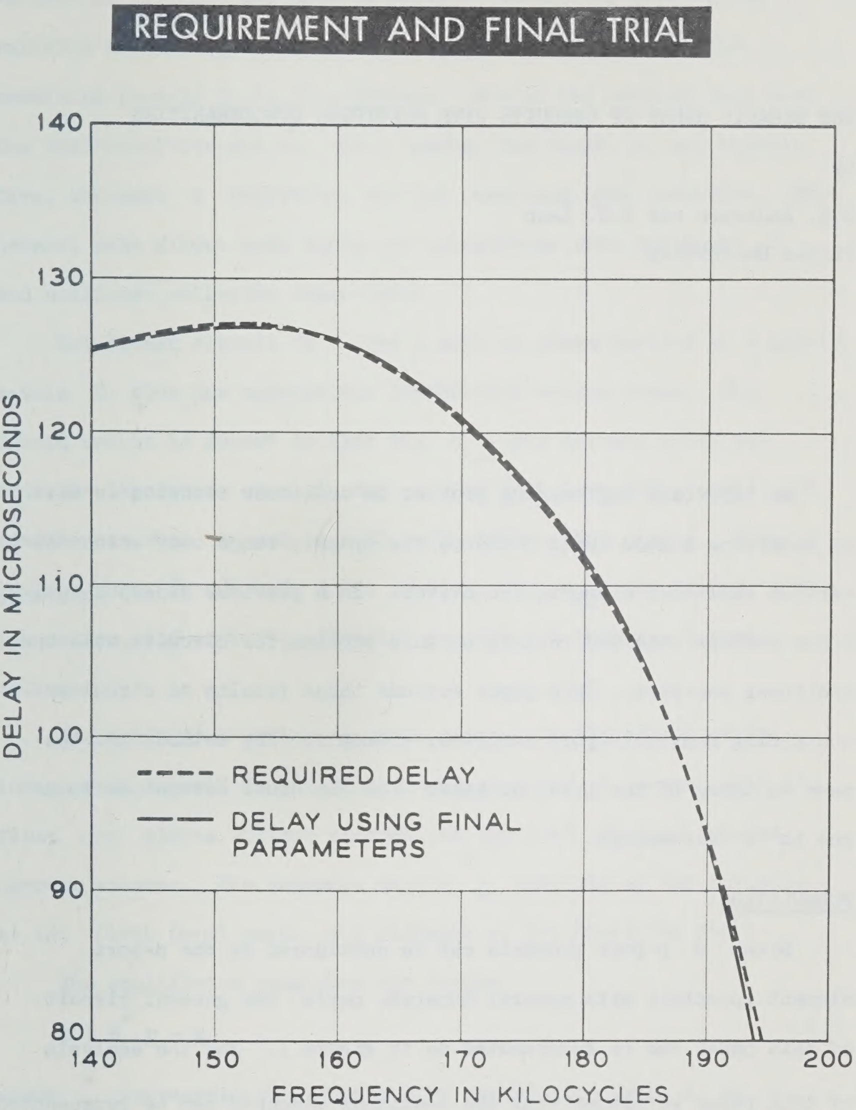
20. Seeking Civilization when Lost in the Woods


OPTIMIZATION OF NETWORK PERFORMANCE
REQUIRE-MENTS
NETWORK DESCRIP-TION
TRIAL PARA-METERS
SUM OF SQUARED DIFFERENCES
ANALYSIS
STOP
CONVER-GENCE CHECK
DIRECTION OF IMPROVE-MENT STEP
SIZE OF IMPROVE-MENT STEP
IMPROVED PARAMETERS

21. Optimization of Network Performance (Flow Chart)



22. Requirement and First Trial



23. Requirement and Final Trial

THE DYNAMIC RANGE OF CIRCUITS WITH MULTIPORT NONLINEARITIES

by

D.R. Anderson and B.J. Leon
Purdue University

An important engineering problem in nonlinear circuits is that of obtaining a good upper bound on the dynamic range over which the various nonlinear elements are driven. In a previous paper, Reference 1, the authors obtained results on this problem for circuits with one nonlinear one-port. This paper extends those results to circuits containing many multiport nonlinear elements. The methods are the same as those of the previous paper with the usual extensions from one to n dimensions.

FORMULATION:

Since m p-port elements can be considered as one n-port element (possibly with several separate parts) the general circuit of this paper can be represented as in Figure 1. For the analysis of this paper we assume that the nonlinear elements can be represented by a vector $\underline{x}$. Since we should like to be able to include capacitive, resistive, and inductive nonlinearities we assume that the first $m+p$ components of $\underline{x}$ are voltages, and the remaining components are currents. The components are functions of the response vector at the various ports. For the first m ports we assume the independent variable is charge, at the next p it is current, and at the remaining (n-m-p) it is flux linkage. Now in the special case when the nonlinearities are all one - ports, the first m are capacitive, the next p resistive, and the remaining ones inductive. The general case allows such things as transistors with nonlinear α and nonlinear collector capacitance.

The linear circuit of Figure 1 will be characterized by a hybrid matrix H plus the appropriate independent source terms. The hybrid matrix is formed so that the h_{ij} are driving point and transfer impedances when i and j are both less than (m+p) and admittances when i and j are both greater than (m+p). When i is less than (m+p) but j is greater, the h_{ij} are voltage transfer ratios, and when the situation is reversed they are current ratios.

The excitation vector $\underline{e}$ that represents the effect of the independent sources in the linear network is an n vector with the first m+p places voltage sources and the last (n-m-p) places current sources. The response vector $\underline{r}$ consists of the currents at the first (m+p) ports and voltages at the remaining ports.

The equilibrium equations now become

$$\underline{x} + H \overset{*}{\underset{a}{}} \underline{r} = \underline{e} \tag{1}$$

where $\overset{*}{\underset{a}{}}$ represents the operation of convolution from a to t; a is an equilibrium point as discussed in Reference 1. Since $\underline{x}$ involves charges in the first m places and fluxes in the last (n-m-p) places, we integrate the second vector by parts to get the same set of variables.

After the integration by parts the response vector, $\underline{r}_1$, will have charges in the first m places, currents in the next p places, and flux linkages in the remaining places. As a vector valued function of a vector, $\underline{x}$ is a function of $\underline{r}_1$. The hybrid matrix will be replaced by a new matrix H_1 derived from the old by differentiating the response functions in the first m and last (n-m-p) columns. In addition to $H_{1a}^{*}\underline{r}_1$ some terms will arise due to evaluation of the uv term in the integration by parts formula at the end points. For these extra terms we define H_o as a matrix of constants $h_{ij}(0)$ for $j \leq m$ and $j > m+p$, and zeros in the p central columns. Similarly H_a will denote the matrix of functions $h_{ij}(t-a)$ for $j \leq m$ and $j > m+p$, and zeros in the p central columns. Finally we use $\underline{r}_a$ for the vector of constants formed by evaluating $\underline{r}_1$ at $t = a$.

With the new notation (1) becomes

$$\underline{x} + H_o\underline{r}_1 + H_{1a}^{*}\underline{r}_1 = \underline{e} - H_a\underline{r}_a . \tag{2}$$

This equation is the vector analog of (3) of Reference 1. We can proceed analogously to the derivation of (7) of that paper by separating the equation into an equilibrium part at $t = a$ and a variable part. To this end we define $\underline{r}_v = \underline{r}_1 - \underline{r}_a$; $\underline{e}_v = \underline{e} - \underline{e}_a$ where $\underline{e}_a$ is the constant vector obtained by evaluating $\underline{e}$ at $t = a$.

The next step is to separate the nonlinear vector function $\underline{x}$ into a linear part plus a nonlinear part that is zero when $\underline{r}_v$ is zero. Since $\underline{x}$ is a vector function of the response vector $\underline{r}_1$, we may apply Taylor's theorem to each function about the members of $\underline{r}_a$. Specifically $\underline{x}(\underline{r}_1) = \underline{x}(\underline{r}_a + \underline{r}_v) = \underline{x}(\underline{r}_a) + X_a\underline{r}_v + \underline{n}(\underline{r}_v)$

where X_a is the matrix of partial derivatives of the components of $\underline{x}$ evaluated at $\underline{r}_a$ and $\underline{n}(\underline{r}_v)$ is the remainder from Taylor's theorem.

Since the initial conditions at $t = a$ were included in $\underline{e}$, $\underline{e}_a$ must be equal to $\underline{x}(\underline{r}_a)$. Thus these two terms cancel each other. Furthermore the $H_a\underline{r}_a$ term in (2) is cancelled by $H_{1a}\overset{*}{}\underline{r}_a$ from the left side when $\underline{r}_v + \underline{r}_a$ is used instead of $\underline{r}_1$. The result is

$$\left[X_a + H_o\right] \underline{r}_v + H_{1a}\overset{*}{}\underline{r}_v = \underline{e}_v - \underline{n}(\underline{r}_v). \tag{3}$$

The left side of (3) is a linear, convolution, operator operating on $\underline{r}_v$. If this operator can be inverted, the equation is in standard form for iteration. Then all the results on dynamic range carry over from Reference 1 if the correct norms are used.

THE GENERALIZED PALEY-WIENER THEOREM:

In the one dimensional case the authors used a theorem of Paley and Wiener to show the stability of the left side of the equivalent of (3). The original theorem, and its generalization to matrices, are as follows:

Theorem (Paley-Wiener Reference 2): Suppose $F(s)$ is the Laplace transform of an absolutely integrable function $f(t)$ such that $f(t) = 0$ for $t < 0$. Then $\frac{F(s)}{1+F(s)}$ is the Laplace transform of an absolutely integrable function $\omega(t)$ for $t \geq 0$ if and only if $\left[1+F(s)\right] \neq 0$ for $\mathrm{Re}\,[s] \geq 0$.

The generalization of the theorem to matrices is:

Given an nxn matrix M of real valued absolutely integrable functions of t all of which vanish for $t < 0$. Let M_s denote the Laplace Transform of M and let I be the nxn identity matrix. Then $\left[I+M_s\right]^{-1}$ exists for $s = j\omega$ and $W_s = \left\{\left[I+M_s\right]^{-1} - I\right\}$ is the Laplace transform of a matrix W with properties the same as those assumed for M if and only if $\left[I+M_s\right]^{-1}$ exists for $Re[s] \geq 0$.

Proof: First we note that sums and convolutions of real-valued, absolutely integrable functions of t vanishing for $t < 0$ yield functions of the same property. Thus the matrix of minors of $\left[I+M_s\right]$ can be written as the sum of I and a matrix of Laplace transforms of real valued absolutely integrable functions of t vanishing for $t < 0$.

Next we note that the assumption that $\left[I+M_s\right]^{-1}$ exists for $Re[s] \geq 0$ is equivalent to $\det\left[I+M_s\right] \neq 0$ for $Re[s] \geq 0$. Moreover $\det\left[I+M_s\right]$ is of the form $1+F(s)$ where $F(s)$ is the Laplace transform of a real-valued, absolutely integrable function of t vanishing for $t < 0$. Thus by the original Paley-Wiener theorem $1/\det\left[I+M_s\right]$ is of the form $1+G(s)$ where $G(s)$ is the Laplace transform of a real-valued, absolutely integrable function of t vanishing for $t < 0$.

The sufficiency conclusion about the matrix W follows immediately. Application of the necessity assertion in the original Paley-Wiener result to $\det\left[I+M_s\right]$ yields our necessity proof.

APPLICATION OF PALEY-WIENER

To apply the above theorem to (3) we first multiply through by $\left[X_a + H_o\right]^{-1}$. When the nonlinearities are all one-ports with positive incremental parameters, this inverse always exists. For multiport nonlinearities one must check each case. To see when the inverse might not exist we should note that the matrix H_o contains the reciprocals of the parasitic shunt capacitances of the linear network in its first mxm corner, zeros in the p central columns and reciprocals of parasitic series inductances in the last block on the diagonal. The remaining off diagonal terms may also be nonzero if there are appropriate parasitic reactance tie sets from the inductive ports to the capacitive ports. A detailed study of the allowable cases can be made along the lines of Reference 3. As in that reference it may be necessary to consider some of the nonlinear resistances as nonlinear conductances and introduce a fourth category in the vector $\underline{x}$. The principle remains the same.

After multiplication by $\left[X_a + H_o\right]^{-1}$ the Laplace transform of the operator on the left side of (3) is $\left\{I + \left[X_a + H_o\right]^{-1} H_{1s}\right\}$. This is of the form required for the application of the generalized Paley-Wiener theorem. Now H_1 was derived from the hybrid matrix H by differentiating certain of the response functions. If the original linear circuit were stable in the sense of every bounded input gives a bounded output, then all the original response functions are absolutely integrable. In the lumped circuit case this stability restriction is sufficient to guarantee that the derivatives of the response functions are also absolutely integrable. In the distributed circuit case we must simply assume that the derivatives of the response function are absolutely integrable. Thus the first hypothesis of the theorem is satisfied if the linear circuit is stable and certain ideal jump behaviour is excluded for the distributed case.

The second hypothesis, that on the real part, can readily be shown when the linear circuit is lumped and passive. Since $\left[X_a + H_o\right]$ is a nonsingular matrix of constants the real part condition can just as easily be applied to $\left[X_a + H_o + H_{1s}\right]$ as to the form that fits the statement of the theorem. From the Laplace transform formula for the derivative

$$H_{1s} = H_s[s] - H_o$$

where $[s]$ is a diagonal matrix having s in the first m places, 1 in the next p, and s in the remaining places. Thus the problem reduces to an investigation of $\left[X_a + H_s[s]\right]$.

To begin our investigation let us examine $\left\{X_a[s]^{-1} + H_s\right\}$. This is the hybrid matrix for a circuit consisting of the original linear circuit in parallel with a linearized version of the nonlinear circuit. If this overall circuit is passive and its impedance matrix is PR then the resulting hybrid matrix is also PR (see Reference 4). From the energy function derivation of the PR conditions we can see that the presence of a j axis zero in the quadratic form shows the existence of a frequency at which the circuit is lossless to inputs at some ports. To satisfy the theorem we must require that the overall circuit presents some loss at all frequencies at each port. We call this condition strictly PR since the PR inequality is now a strict inequality.

In the case where the nonlinear element is a set of one ports with positive incremental parameters, then all we need to require is that H_s itself be strictly PR. In this case the addition of the linearized nonlinear circuit merely adds positive elements on the diagonal of H_s and does not disturb the PR character. When coupling elements, especially elements with gain, are in the nonlinear network, each case must be examined separately.

When the hybrid matrix $\left\{X_a[s]^{-1} + H_s\right\}$ is strictly PR, then the matrix $\left\{X_a + H_s\,[s]\right\}$ satisfies the hypotheses of the theorem if

X_a is nonsingular. The determinant of the second matrix, the matrix to be tested for the theorem, is just $(\frac{1}{s})^{n-p}\det\left\{X_a[s]^{-1}+H_s\right\}$. If the brace is a strictly PR matrix the result is clearly nonzero for $s \neq 0$. When $s = 0$, the determinant in question is just $\det\left[X_a+\hat{H}_s\right]$ where $\hat{H}_s$ is zero in the first m columns and in the last n-m-p columns and agrees with H_s at $s = 0$ in the remaining columns. Again if the nonlinearity consists of all positive one-ports, there is no question of the application of the theorem.

In summary then we can say that if X_a and $\left[X_a + H_o\right]$ are both nonsingular and if the hybrid circuit matrix $\left\{X_a[s]^{-1}+H_s\right\}$ is strictly PR, then in the lumped circuit case the Paley-Wiener theorem applies and (3) can be written

$$\underline{r}_v = \left[X_a + H_o\right]^{-1}\left\{\left[\Lambda + \hat{H}_1\right]\left[\underline{e}_v - \underline{n}(\underline{r}_v)\right]\right\} \qquad (4)$$

where

$$\hat{H}_{1s} = \left\{\left[I + \left[X_a + H_o\right]^{-1} H_{1s}\right]^{-1} - I\right\}$$

Λ is the identity convolution operator, a diagonal matrix of δ functions.

THE DYNAMIC RANGE RESULT

Equation (4) is the matrix analog of equation (11) of Reference 1. The same iteration scheme applies if the sup of functions in the reference is replaced by the maximum of the sups of the components of the vector function. For our matrix operators the norm must be the max over all rows of the sum of the integrals of the absolute values of the operators in that row.

The final dynamic range result is

$$|| \underline{r}_v || \leq \frac{|| \underline{r}_e ||}{1 - d} \tag{5}$$

where $\underline{r}_e$ is the solution to the completely linearized circuit when

$\underline{e}_v$ is the excitation.

d is the smallest solution of the equation

$$\xi = || \left[X_a + H_o \right]^{-1} \left\{ \Delta + \hat{H}_1 \right\} || \; || \underline{n}' \left(\frac{1}{1-\xi} \right) || \underline{r}_e ||) || \tag{6}$$

where $\underline{n}'$ is a monotonic vector function whose components are

$$n_i'(y) = \sup_{-y \leq \eta_i \leq y} \left[\sum_{j=1}^{n} \left| \frac{\partial n_i}{\partial \eta_j} (\underline{\eta}) \right| \right]$$

Now (5) is the analog of (16) of Reference 1 and (6) the analog of (22). The fact that (6) has a solution less than one if e_v is made sufficiently small follows by the same argument as that used for (22) in the reference.

CONCLUSIONS

The dynamic range of Reference 1 does indeed generalize to the problem of m - port nonlinearities if the functional form of the nonlinearity is restricted so that there is only one independent variable associated with each port and, as in the one dimensional case, the nonlinear characteristics are twice differentiable. The restrictions on the linear circuit are those imposed by the Paley-Wiener theorem -- just as in the one dimensional case. When the n-port nonlinearity consists of n one-port with positive incremental parameters, the method applies so long as the linear network is lumped and strictly PR. The result very likely carries over to distributed linear networks, but to prove this would require generalizing the known results on hybrid matrices to the distributed cases. Finally the results further generalized to nonlinearities with coupling provided that the gain of the nonlinear element is not such that it upsets the PR character of the overall linearized circuit at the initial condition point.

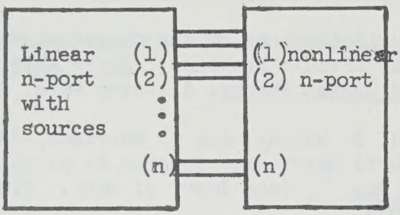

Linear
n-port
with
sources
(1)
(2)
(n)
(1) nonlinear
(2) n-port
(n)

Figure 1. General Circuit

BIBLIOGRAPHY

1. B.J. Leon and D.R. Anderson, "The Stability of Pumped Nonlinear Reactance Circuits," IEEE Trans. CT-10, pp. 468-476, Dec. 1963.

2. REAC Paley and N. Wiener, "Fourier Transforms in the Complex Domain," American Math. Soc. Colloquium Pub., New York, vol. 19, sect. 19, Th. XVIII, pp. 60-61; 1934.

3. B.J. Leon, "Limitations on the Classical Volterra Equation Theory in Nonlinear Circuit Problems," First Allerton Conference on Circuit and System Theory, Allerton Park, Ill., 1963.

4. R.J. Duffin, D. Hazony, and N. Morrison, "Network Synthesis Through Hybrid Matrices," Theorem 2, p. 9, AFCRL63-568, Scientific Rep. 4, Case Inst. of Tech., Cleveland, Ohio, Aug. 18, 1964.

ON SYSTEM FUNCTIONS WITH THE PROPERTY OF SEPARABILITY*

L. Weiss
Divisions of Applied Mathematics and Engineering
Brown University
Providence, R. I.

1. INTRODUCTION

The system function concept, as expounded by Zadeh [1], among others, has proved moderately useful for representing linear differential systems in a number of situations. Of particular interest are those systems whose input/output behavior can be described by system functions which possess the property of separability, for those systems are particularly easy to analyze and synthesize. The property of separability is defined for a general function as follows.

Definition: A scalar function H of two variables p,q is said to be separable in those variables if it can be written in the form

$$H(p,q) = \sum_{i=1}^{N} \alpha_i(p)\beta_i(q)$$

where each α_i is an ordinary function of p and each β_i is an ordinary function of q. In this paper we investigate some ramifications of the property of separability in a particular class of system functions.

---

* This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research under Grant AF-AFOSR-693-64.

2. CLASS OF SYSTEMS TO BE CONSIDERED

We shall be concerned only with single-input, single-output nonanticipative (i.e. causal) systems which can be described by a differential equation of the following type.

$$(1)\qquad a_n(t)\frac{d^n y}{dt^n} + a_{n-1}(t)\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_o(t)y$$

$$= b_m(t)\frac{d^m u}{dt^m} + \cdots + b_o(t)u \ , \ m < n$$

where

u = input (scalar)
y = output (scalar)
$a_n(t) \neq 0$ for any t
m,n are minimal.

The representation (1) implies and is implied by a representation of the form

$$(2)\qquad \frac{dx}{dt} = F(t)x + g(t)u(t)$$

$$y(t) = h(t)x(t)$$

where

x = n-vector (the state vector)
$F(t)$ = nxn matrix
$g(t)$ = n-vector (column)
$h(t)$ = n-vector (row)
$F(\cdot)$, $g(\cdot)$, $h(\cdot)$ are real and are assumed to be differentiable as many times as needed.

This is the so-called state variable representation of the system. It is not necessary, in what follows, to know exactly what $F(\cdot)$, $g(\cdot)$, $h(\cdot)$ are. All we need to know is that (2) is valid.

3. THE IMPULSE RESPONSE

The solution of (2) for zero initial state at time $t=0$ is given by [2]

$$y(t) = \int_{0}^{t} h(t)\Phi(t,\tau)g(\tau)u(\tau)d\tau \tag{3}$$

where

$\Phi(t,\tau)$ = nxn fundamental matrix solution to the homogeneous version of (2).

That is, $\Phi(t,\tau)$ satisfies the relation

$$\frac{d\Phi}{dt} = F(t)\Phi , \quad \Phi(t,t) = I \tag{4}$$

where I is the nxn identity matrix. The matrix Φ also has the so-called semi-group property given below.

$$\Phi(t,\tau) = \Phi(t,\xi)\Phi(\xi,\tau) \quad \text{for all } t,\tau,\xi . \tag{5}$$

If we replace u by a δ-function in (3), then from the well known properties of δ-functions we obtain the result that the impulse response of the system $W(t,\tau)$ is given by

$$\begin{aligned} W(t,\tau) &= h(t)\Phi(t,\tau)g(\tau) , \quad t \geq \tau \\ &= 0 \qquad\qquad\qquad\quad , \quad t < \tau . \end{aligned} \tag{6}$$

The last part of (6) arises from the nonanticipative or causal nature of the system. As long as the system is asymptotically stable, the input/output relation can be written as

$$y(t) = \int_{-\infty}^{\infty} W(t,\tau)u(\tau)d\tau . \tag{7}$$

As far as this paper is concerned, the most important property contained by the system's impulse response is given by:

LEMMA 1: A scalar function W of two variables t,τ, which vanishes for $t < \tau$ can be regarded as an impulse response for a system (1), (2) if and only if it is a separable function of the variables t,τ, i.e., if and only if it can be written in the form

$$(8) \qquad W(t,\tau) = \Psi(t)\Theta(\tau) = \sum_{i=1}^{n} \varphi_i(t)\theta_i(\tau), \quad t \geq \tau$$
$$= 0 \qquad t < \tau .$$

($\Psi(\cdot)$ is a row vector with components $\{\Psi_i\}$ and Θ is a column vector with components $\{\theta_i\}$)

Proof (sufficiency): In (2) set $h = \Psi$, $g = \Theta$.

(necessity): For any t_o, a given impulse response (8) can be written as (according to (5))

$$W(t,\tau) = h(t)\Phi(t,t_o)\Phi(t_o,\tau)g(\tau), \quad t \geq \tau$$
$$= 0 \qquad , t < \tau .$$

Now set

$$\Psi(t) = h(t)\Phi(t,t_o) , \ \Theta(\tau) = \Phi(t_o,\tau)g(\tau). \qquad \text{Q.E.D.}$$

This lemma will play an important role later on. It is interesting to note that the separability property of a system's impulse response in the time domain implies the same property for the response of the system to an impulse function in a transform domain (see Appendix).

4. THE T-TRANSFORM

The reader is referred to [5] for a general discussion of the system function concept. We shall assume the necessary general background and merely present a brief account of those definitions and formulas which are needed for the remainder of the paper.

A system function is defined in conjunction with an integral transform. Unless otherwise specified, the integral transform dealt with in this paper is the T-transform [6] defined as follows. For a suitably chosen [6] time function $f(\cdot)$, we have the transform pair

$$(9) \qquad f(t) = \int_{c_k} F(\lambda)k(t,\lambda)d\lambda = T^{-1}_{(\lambda;t)}\{F(\lambda)\}$$

$$(10) \qquad F(\lambda) = \int_{-\infty}^{\infty} f(t)\ell(\lambda,t)dt = T_{(t;\lambda)}\{f(t)\}$$

where

$$(11) \quad k(t,\lambda) = \text{inverse transform kernel}$$

$$= \frac{1}{2\pi i} \exp\left\{-\int_0^t \frac{b(\xi)-\lambda}{a(\xi)}\, d\xi\right\}$$

$$(12) \quad \ell(\lambda,t) = \text{transform kernel} = \frac{1}{a(t)} \exp\left\{\int_0^t \frac{b(\xi)-\lambda}{a(\xi)}\, d\xi\right\}$$

$a(\cdot)$, $b(\cdot)$ are C^{∞} functions and $a(t) \neq 0$ for any t

c_k = integration path associated with inverse transform (a closed contour enclosing all poles of integrand)

and a subscript $(p;q)$ on the symbols T or T^{-1} means "integration taken with respect to p, with q a transform parameter."

It is easily demonstrated that $k(t,\lambda)$ and $\ell(\lambda,t)$ satisfy the following equations respectively.

$$(13) \qquad [a(t)\frac{d}{dt} + b(t)]k(t,\lambda) = \lambda k(t,\lambda)$$

$$(14) \qquad [-\frac{d}{dt}(a(t)\cdot) + b(t)]\ell(\lambda,t) = \lambda\,\ell(\lambda,t)$$

Equation (14) is recognized as the formal adjoint [4] of (13).

It should be noted that the T-transform includes, as special cases, the Laplace, Fourier, and various versions of the Mellin transform. Our results will therefore take on some measure of generality.

5. SYSTEM FUNCTIONS BASED ON THE T-TRANSFORM

We shall consider two system functions, $H(\lambda,t)$ and $\hat{H}(\lambda,t)$, defined in conjunction with the T-transform.

These system functions can be expressed, for a particular system, as below.*

$$(15) \qquad H(\lambda,t) = \left.\frac{\text{output}}{\text{input}}\right|_{\text{input}=k(t,\lambda)}$$

$$(16) \qquad \hat{H}(\lambda,t) = \left.\frac{\{T \text{ output}\}}{\{T \text{ input}\}}\right|_{T\{\text{input}\}=\ell(\lambda,t)}$$

(There are alternate definitions for these system functions and an interesting relationship between them which involves the concept of dual dynamical systems. This is discussed in another paper [5].)

It follows from (7), (15), and (16) that the relationship between the system functions and the system's impulse response is given by

$$(17) \qquad H(\lambda,t)k(t,\lambda) = \int_{-\infty}^{\infty} W(t,\tau)k(\tau,\lambda)d\tau$$

$$(18) \qquad \hat{H}(\lambda,t)\ell(\lambda,t) = \int_{-\infty}^{\infty} W(\tau,t)\ell(\lambda,\tau)d\tau \, .$$

We note in passing that, for a given system, if H turns out to be a function only of λ, then H and $\hat{H}$ are identical and k(t,λ) is referred to as a characteristic input for that system.

The properties of H and $\hat{H}$ which motivate their study concern the role played by these system functions in input/output expressions when initial conditions are zero. Under such conditions one can show that the time-domain output y(t) of a given system, corresponding to input u(t), can be expressed by the relation

---

* For the case where T is the Laplace or Fourier transform operator, the system function (15) was introduced by Zadeh [1] and the system function (16) by Gersho [2] and (independently) Bello [3].

$$(19)\qquad y(t) = T^{-1}_{(\lambda;t)}\{H(\lambda,\xi)U(\lambda)\}\Big|_{\xi=t}$$

where

$$(20)\qquad U(\lambda) = T_{(t;\lambda)}\{u(t)\}.$$

In like manner, the T-transform-domain output $Y(\lambda)$ of the system, corresponding to input $u(t)$, is given by

$$(21)\quad Y(\lambda) = T_{(t;\lambda)}\{y(t)\} = T_{(t;\lambda)}\{\hat{H}(\lambda',t)u(t)\}\Big|_{\lambda'=\lambda} .$$

6. SEPARABLE SYSTEM FUNCTIONS

From the definition of separability of a function of two variables, we see that if $H(\lambda,t)$ is separable in λ and t, it can be written in the form

$$(22)\qquad H(\lambda,t) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(\lambda) .$$

Such system functions have been discussed in the literature for the case where T is the Laplace or Fourier transform operator (see [7]). The interest in such system functions stems from the physical interpretation they immediately suggest. If, for a particular system, $H(\lambda,t)$ is given by (22), then the time-domain output of the system (for zero initial conditions) for an input u becomes

$$(23)\qquad y(t) = \sum_{i=1}^{N} \alpha_i(t)T^{-1}_{(\lambda;t)}\{\Phi_i(\lambda)U(\lambda)\}$$

where U is the T-transformed input function. Equation (23) indicates that the input/output relation can be realized by N parallelled systems each of which consists of a nonzero-memory filter with characteristic input $k(t,\lambda)$ followed by a time-varying-gain amplifier as in Figure 1.

In like manner, suppose that, for a given system, the system function $\hat{H}(\lambda,t)$ is separable, e.g.,

$$\hat{H}(\lambda,t) = \sum_{i=1}^{N} \beta_i(t)\Psi_i(\lambda). \tag{24}$$

Then the T-transform domain output $Y(\lambda)$ for a given input u is given by

$$Y(\lambda) = \sum_{i=1}^{N} \Psi_i(\lambda)T_{(t;\lambda)}\{\beta_i(t)u(t)\} . \tag{25}$$

It follows from (25) that the input/output relation can be realized by N parallelled systems each of which consists of a time-varying-gain amplifier followed by a nonzero-memory filter with characteristic input $k(t,\lambda)$ as in Figure 2.

If it turns out that both H and $\hat{H}$ are separable for a given system, then that system can be realized in either of the ways just discussed. The actual choice might be made, for instance, on the basis of stability considerations [8].

7. THEOREM ON SEPARABLE SYSTEM FUNCTIONS

We now present a theorem which relates separability of a system function in t and λ to rationality of the system function in λ. The result holds for both H and $\hat{H}$. A special case of this theorem was originally given by Cruz [7] and (independently) Kaplan [9] for the system function H. The proof given here differs from those in [7] and [9].

THEOREM 1. The system functions $H(\lambda,t)$, $\hat{H}(\lambda,t)$ associated with the T-transform are separable in t and λ if they are rational in λ.

Proof. The proof follows the same pattern for both H and $\hat{H}$. We therefore present the proof only for $\hat{H}$ and leave the remainder as an exercise.

If $\hat{H}(\lambda,t)$ is rational in λ, then, assuming no multiple poles,* it can be written as

---

* This causes no loss of generality in the result, but simplifies the notation considerably.

(26) $$\hat{H}(\lambda,t) = \sum_{i=1}^{N} \frac{\beta_i(t)}{\lambda+\lambda_i(t)} .$$

From (18) and the definition of the T-transform it follows that

(27) $$W(\tau,t) = \int_{c_k} \hat{H}(\lambda,\tau)\ell(\lambda,\tau)k(t,\lambda)d\lambda .$$

Substituting the expression for each integrand factor, we have

$$W(\tau,t) =$$

$$\sum_{i=1}^{N} \frac{\beta_i(\tau)}{a(t)} \exp[-\int_{\tau}^{t} \frac{b(\xi)}{a(\xi)} d\xi \int_{c_k} \frac{1}{\lambda+\lambda_i(\tau)} \exp[\lambda\int_{\tau}^{t} \frac{d\xi}{a(\xi)}]\frac{d\lambda}{2\pi i}$$

which, by the definition of the contour c_k, becomes

$$W(\tau,t) = \sum_{i=1}^{N} \frac{\beta_i(\tau)}{a(\tau)} \exp -\int_{\tau}^{t} \frac{b(\xi)}{a(\xi)} d\xi]\exp[-\lambda_i(\tau)\int_{\tau}^{t} \frac{d\xi}{a(\xi)}] .$$

Now, since $W(\tau,t)$ must be separable in τ and t (see Lemma 1), then each term

$$\exp[-\lambda_i(\tau)\int_{\tau}^{t} \frac{d\xi}{a(\xi)}]$$

must be separable in τ and t. This is clearly possible only when

$$\lambda_i = \text{constant}. \qquad \text{Q.E.D.}$$

COROLLARY. If a given system function H or $\hat{H}$ is rational in λ but is not separable in t and λ, then neither H nor $\hat{H}$ can be associated with a system described by an ordinary differential equation.

Proof. Nonseparability of any system function that is rational in λ implies nonseparability of an impulse response which, by the Lemma is not commensurate with a linear differential system. Q.E.D.

Sufficient conditions on the impulse response of a system to ensure separability of the system functions H and/or $\hat{H}$ are provided by the following.

THEOREM 2. *Let the impulse response of a given system be*

$$(28)\qquad W(t,\tau) = \sum_{i=1}^{n} \psi_i(t)\theta_i(\tau), \quad t \geq \tau$$

$$= 0 \qquad\qquad , \quad t < \tau.$$

Then the system function $H(\lambda,t)$ *will be rational in* λ *if each function* $\theta_i(t)$ *is of the form* $\sum_j \alpha_{ij}\ell(\lambda_{ij},t)$, *where the* $\{\alpha_{ij}\}$ *and* $\{\lambda_{ij}\}$ *are constants.* *Similarly, the system function* $\hat{H}(\lambda,t)$ *will be rational in* λ *if each function* $\psi_i(t)$ *is of the form* $\sum_j \beta_{ij}k(t,\eta_{ij})$ *where the* $\{\beta_{ij}\}$ *and* $\{\eta_{ij}\}$ *are constants.**

Proof. Follows by making the indicated substitutions for the $\{\psi_i\}$ and $\{\theta_i\}$ in (17) and (18).

8. THE APPROXIMATION PROBLEM

The advantages of having a separable system function representation of a system is sufficient motivation for investigating the possibility of approximating systems described by nonseparable functions by systems whose system functions are separable. For simplicity, our attention is restricted only to system functions based on the Laplace transform.

The approach to the solution to the approximation problem is immediately suggested by the following Corollary to Theorem 2.

---

* The conditions on $\{\theta_i\}$ and $\{\psi_i\}$ are necessary as well as sufficient if there are no multiple poles in the partial fraction expansions of H and $\hat{H}$ respectively.

COROLLARY. If T is the Laplace transform operator then a system with impulse response (28) can be described by a separable system function $H(\lambda,t)$ $[\hat{H}(\lambda,t)]$ if and only if the functions $\{\Theta_i\}[\{\psi_i\}]$ form an exponential basis.

The problem of finding an exponential basis to represent a set of functions given either in numerical or closed form is a well known problem in signal theory, and many aspects of the problem have been completely solved. We shall not discuss this question here but refer the reader to [10], [11].

Consider, however, the case where the functions Θ_i and ψ_i are not explicitly given either numerically or analytically, but the impulse response $W(t,\tau)$ is available in the form of regularly spaced points of data on the t,τ plane. Then, assuming the system to be asymptotically stable so that the data can be truncated in the direction of increasing t and decreasing τ, we can in principle use Prony's classical method [12] (or the more efficient Weiss-McDonough version of it [13]) to obtain $W(t,\tau)$ as a summation of products of exponentials in t with exponentials in τ such that the resulting function fits the truncated set of data points exactly. The procedure would be to use the Prony-Weiss-McDonough method to find the exponentials which fit the data along the τ axis for each value of the discrete variable t. Then treat the coefficients of the aggregate of τ-exponentials as discrete values of some function which in turn can be expressed as a sum of exponentials, using the aforementioned method. In this way we obtain a representation for $W(t,\tau)$ corresponding to which both system functions H and $\hat{H}$ are separable.

REFERENCES

1. L.A. Zadeh, "Frequency Analysis of Variable Networks", Proc. I.R.E., Vol. 38, No. 3, pp.291-299; March 1950.
2. A. Gersho, "Characterization of Time-Varying Linear Systems", Proc. I.E.E.E., Vol. 51, No. 1, p.238; 1963.
3. P. Bello, "Time-Frequency Duality", I.E.E.E. Trans. on Information Theory, Vol. IT-10, No. 1; 1964.
4. E.A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, New York; 1955.
5. L. Weiss, "Dual Dynamical Systems and Their Representation by Systems Function," these Proceedings.
6. L. Weiss, "On the Foundations of Transfer Function Analysis", Int'l. J. Engrg. Sci. (to appear).
7. J.B. Cruz, Jr. and M.E. Van Valkenburg, "The Synthesis of Models for Time-Varying Linear Systems", Proc. P.I.B. Symp. on Active Networks, April, 1960.
8. L. Weiss, "The Role of System Functions in the Representation and Analysis of Linear Systems", RIAS Tech. Rept. TR-64-4; April, 1964.
9. W. Kaplan, Operational Methods for Linear Systems, Addison-Wesley, Reading, Mass.; 1962.
10. W.H. Huggins, "Representation and Analysis of Signals, Part I. The Use of Orthogonalized Exponentials", (Report) The Johns Hopkins University, Baltimore, Md.; September, 1957.
11. R.N. McDonough, "Representation and Analysis of Signals, Part XV. Matched Exponents for the Representation of Signals", (Report) The Johns Hopkins University, Baltimore, Md.; April, 1963.

12. R. Prony, "Essai expérimental et analytique, etc.", Paris, J. l'Ecole Polytechnique, 1, cahier 2, pp.24-76; 1795.

13. L. Weiss and R.N. McDonough, "Prony's Method, Z-transforms, and Padé Approximation", S.I.A.M. Rev., Vol. 5, No. 2, pp.145-149; April, 1963.

APPENDIX

Transform-Domain Impulse Responses*

Suppose we are interested only in T-transformed input/output relations (see Section 4). Then, just as an impulse response provides a convenient means for representing time-domain input/output relations, there is a function of two variables which plays an analogous role in the T-transform domain. To find the latter for a given system consider the following.

Let $\delta(\lambda-\eta)$ be a δ-function in the T-transform domain. This δ-function is formally related to the time-domain δ-function $\delta(t-\tau)$ by the expression

$$\text{(A1)} \qquad \delta(\lambda-\eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(t-\tau)\ell(\lambda;t)k(\tau,\eta)\,dt\,d\tau$$

$$= Q\{\delta(t-\tau)\}$$

where Q is an operator defined by (A1). Assuming the impulse response W of the given system to be in the domain of Q, define the function Γ by the relation

$$\text{(A2)} \qquad \Gamma(\lambda,\eta) = Q\{W(t,\tau)\}.$$

If we consider $\delta(\lambda-\eta)$ as a transform-domain impulse input to the system, then the transform domain response to this input is the function $\Gamma(\lambda,\eta)$, which can therefore be referred to as a transform-domain impulse response.

It is readily demonstrated [8] that the T-transformed output Y of a system, corresponding to a T-transformed input U, is given by

---

* We consider only the T-transform domain (see Section 4) in this discussion. However, the results hold for any transform of conventional type, i.e., any transform with one transform kernel and one inverse transform kernel.

$$Y(\lambda) = \int_{c_k} \Gamma(\lambda,\eta)U(\eta)d\eta \ . \tag{A3}$$

The function Γ was discussed by Zadeh [1] for the special case where T is the Laplace or Fourier transform operator. Zadeh referred to it as the "bi-frequency function".

We now present an interesting property of the function Γ which follows immediately from Theorem 1 and (A2).

THEOREM 3. A scalar function Γ of two variables λ,η can be regarded as a transform-domain impulse response for a system which, in the time-domain, is a linear differential system, if and only if Γ is separable in λ and η, i.e. if and only if Γ can be expressed in the form

$$\Gamma(\lambda,\eta) = \sum_{i=1}^{N} \varphi_i(\lambda)\beta_i(\eta).$$

(For a multi-input, multi-output system, the theorem still holds with Γ represented by a product of two matrices one of which is a function λ, the other a function of η.)

$$H(\lambda,t) = \sum_{i=1}^{N} a_i(t)\Phi_i(\lambda)$$

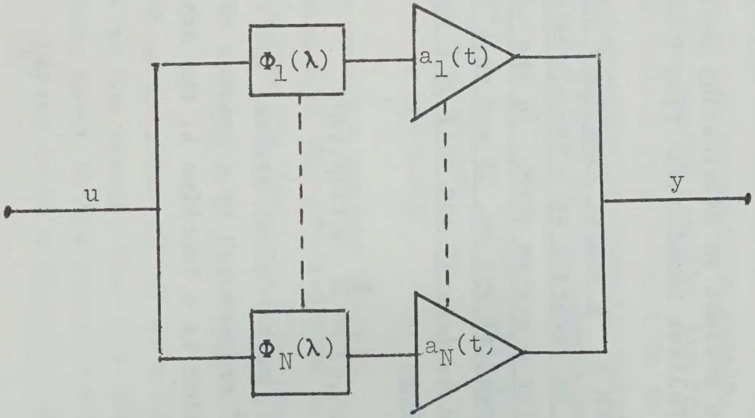

u
$\Phi_1(\lambda)$
$\Phi_N(\lambda)$
$a_1(t)$
$a_N(t)$
y

Figure 1

$$\hat{H}(\lambda,t) = \sum_{i=1}^{N} b_i(t)\Psi_i(\lambda)$$

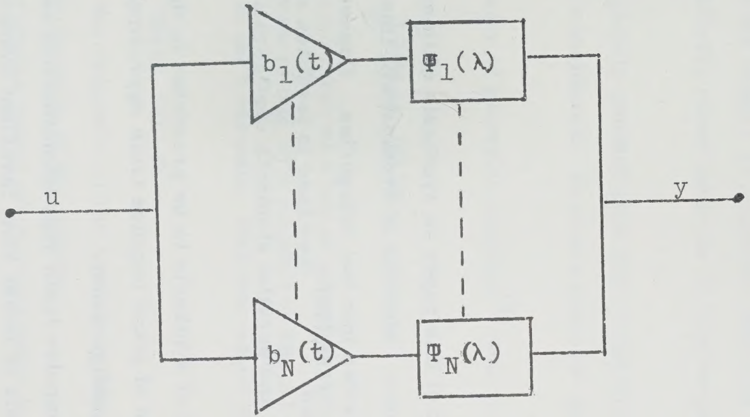

u
$b_1(t)$
$\Psi_1(\lambda)$
$b_N(t)$
$\Psi_N(\lambda)$
y

Figure 2

SYNTHESIS OF LINEAR TIME-VARYING SYSTEMS*

G. S. Glinski and M. Arozullah
Department of Electrical Engineering
The University of Ottawa
Ottawa, Canada

Introduction:

The subject of this paper is synthesis of linear time-varying systems. Methods of synthesis of linear time-varying systems may be divided into two categories. Depending on the type of configuration used(15),(16) they lead to feedback configuration(5),(6),(7),(8),(9),(10) and parallel elements configuration.(3),(11),(12),(13),(14)

The method of synthesis to be presented in this paper, namely, the method of plane impulse train approximation, leads to parallel elements configuration.

Method of impulse train approximation has been used in the past for the synthesis of linear time invariant systems.(17) Cruz(14) extended and applied this method to the problem of synthesis of linear time-varying systems in the time domain from the impulse response $h_1(t,\tau)$.

---

* The research reported herein was supported in part by the National Research Council of Canada under grant no. A-875.

This paper considers the problem of synthesis of such systems by this method from other system functions such as

1. Impulse response $h_2(y,t)$.
2. H-system function $H(\nu,t)$.
3. G-system function $G(\omega,\tau)$.
4. Bifrequency system function $\Gamma(\mu,\nu)$.

Before going into the details of synthesis methods it seems expedient to discuss briefly some methods of representation of such systems.

Only completely controllable and completely observable (5), (18) systems are considered. Systems are also assumed to be stable.

The property of linearity implies (1) that if the input can be expressed as

$$u(\eta) = \int_c k(\eta,\lambda)\ U(\lambda)d\lambda \qquad (1)$$

where $k(\eta,\lambda)$ is a member of a set of possible component functions, $U(\lambda)$ is the spectrum of $u(\eta)$ with respect to $k(\eta,\lambda)$ and c is a suitable contour in the λ-domain, then the output may be expressed as

$$v(\eta) = \int_c K(\eta,\lambda)\ U(\lambda)\ d\lambda \qquad (2)$$

where $K(\eta,\lambda)$ is the response of the system to $k(\eta,\lambda)$ and is called the characteristic function of the system.

The following particular cases of characteristic functions are considered:

1. Time domain specifications (2)

$k(\eta,\lambda) \rightarrow \delta(t-\tau)$; $\lambda \rightarrow \tau$, time of the application of the input

$\eta \rightarrow t$, the time of measurement of the output.

= delayed unit impulse.

$K(\eta,\lambda) \rightarrow h_1(t,\tau)$

= the impulse response of the system.

In terms of the impulse response $h_1(t,\tau)$ the response of the system to an input u(t) is given by

$$v(t) = \int_{-\infty}^{t} h_1(t,\tau)\ u(\tau)d\tau. \qquad (3)$$

Another form of the impulse response is $h_2(y,t)^{(3)}$, where $y = t-\tau$ is the age of the input.

In terms of $h_2(y,t)$ the output of the system is given by

$$v(t) = \int_{0}^{\infty} h_2(y,t)\ u(t-y)dy. \qquad (4)$$

2. Mixed time-frequency domain specification[(2)].

(a) $k(\eta,\lambda) \rightarrow \frac{e^{st}}{2\pi j}$; $\lambda \rightarrow t$

$\eta \rightarrow s$, the complex input frequency.

$K(\eta,\lambda) \rightarrow H(s,t)\ \frac{e^{st}}{2\pi j}$

H(s, t) is called the H-system function and in the particular case of $s = j2\pi\nu$, $H(\nu, t)$ is called the frequency response function of the system.

In terms of H(s, t) the output is given by

$$v(t) = \frac{1}{2\pi j} \int_{c} H(s,t)\ U(s)e^{st} ds \qquad (5)$$

where U(s) is the spectrum of u(t) and c is a suitable contour in the s-domain.

(b) $k(\eta,\lambda) \rightarrow e^{-m\tau}$; $\lambda \rightarrow \tau$,

$\eta \rightarrow m$, the complex output frequency.

$K(\eta,\lambda) \rightarrow G(m,\tau)$

$G(m,\tau)$ is called the G-system function[(4)] of the system.

In terms of $G(m,\tau)$ the spectrum of the output of the system is given by

$$V(m) = \int_{-\infty}^{\infty} G(m,\tau)\ u(\tau)\ e^{-m\tau} d\tau. \qquad (6)$$

3. Frequency domain specification(2).

$k(\eta, \lambda) \longrightarrow \delta(n-s)$; $\eta \rightarrow n$, the complex frequency of system variation.

$\lambda \longrightarrow s$.

$K(\eta, \lambda) \longrightarrow \Gamma(n, s)$

$\Gamma(n, s)$ is called the Bifrequency system function.

In terms of $\Gamma(n, s)$ the spectrum of the output of the system is given by

$$V(n) = \int_c \Gamma(n, s)\ U(s)ds. \tag{7}$$

This concludes the consideration of representation of such systems.

Synthesis by means of Plane Impulse Train Approximation:

A function $f(x, y)$ can be expressed as an integral

$$f(x, y) = \int_{-\infty}^{\infty} f(x, \xi)\quad \delta(y-\xi)d\xi \tag{8}$$

$$= \int_{-\infty}^{\infty} f(y, \xi)\quad \delta(x-\xi)d\xi. \tag{9}$$

The integral may be approximated by the sum:

$$f(x, y) = \int_{-\infty}^{\infty} f(x, \xi)\quad \delta(y-\xi)d\xi$$

$$\cong \sum_{k=1}^{N} \Delta\xi_k \quad f(x, \xi_k) \quad \delta(y-\xi_k)$$

$$= f^*(x, y) \tag{10}$$

where N is a finite integer and $\Delta\xi_k = \xi_{k+1} - \xi_k$.

Equation (10) states that the function $f(x, y)$ may be approximated by an appropriately weighted sum of a train of delayed plane impulses. The plane impulse occurring at $y = \xi_k$ has an area $\Delta\xi_k f(x, \xi_k)$ where $f(x, \xi_k)$ is the value of the function $f(x, y)$ at $y = \xi_k$ and $\Delta\xi_k$ is the distance in the y-direction between the plane impulses occurring at $y = \xi_k$ and $y = \xi_{k+1}$.

Equation (10) forms the basis of the method of synthesis presented here.

1. Model from $h_2(y,t)$.

(i) $h_2(y,t)$ may be approximated by

$$h^*_2(y,t) = \sum_{k=1}^{N} h_2(y,\xi_k)\,\Delta\xi_k\,\delta(t-\xi_k). \tag{11}$$

In terms of this, the approximate output is given by

$$v^*(t) = \int_0^\infty h^*_2(y,t)\,u(t-y)dy$$

$$= \sum_{k=1}^{N} \Delta\xi_k\,v_k(t)\,\delta(t-\xi_k) \tag{12}$$

where $$v_k(t) = \int_0^\infty h_2(y,\xi_k)\,u(t-y)dy. \tag{13}$$

From equations (12) and (13) it is seen that the approximate output $v^*(t)$ may be obtained by

(1) filtering the input through fixed linear systems with impulse responses $h_2(y,\xi_k)$'s.

(2) multiplying the output of the k-th filter by an impulse of area $\Delta\xi_k$ appearing at $t = \xi_k$.

and (3) adding all the resultants.

The corresponding system may be realized in the configuration shown in figure 1. The advantage of this model is that it consists of N time-invariant systems, a sequential switch and a hold circuit and thus does not require any complex components such as multipliers and function generators.

The switch is adjusted in such a way that at $t = \xi_k$ it connects the hold circuit with the output of the k-th filter and the hold circuit maintains this value at the output for an interval $\Delta\xi_k$ until the switch connects the next filter.

The time invariant system functions may not be, in some cases, exactly realizable, but it is assumed that in such cases they may be approximated by other system functions which are realizable and closely equivalent to the original ones.

(ii) $h_2(y,t)$ may also be approximated by

$$h_2^*(y,t) = \sum_{k=1}^{N} h_2(\xi_k,t)\ \Delta\xi_k\ \delta(y-\xi_k). \qquad (14)$$

In terms of this, the approximate output is

$$v^*(t) = \sum_{k=1}^{N} b_k(t)\ u(t-\xi_k) \qquad (15)$$

where $$b_k(t) = \Delta\xi_k\ h_2(\xi_k,t). \qquad (16)$$

The corresponding realization of the system is shown in figure 2. This model consists of a tapped delay line, time-varying gain amplifiers and an adder, and is the same as the one obtained by Cruz.(14)

The time-varying gain amplifiers may be realized by means of multipliers and function generators or amplitude modulators or sequentially gated summing amplifiers.

2. Model from H-system function $H(\nu,t)$ where ν is the input frequency in cycles per second.

(i) $H(\nu,t)$ may be approximated by

$$H^*(\nu,t) = \sum_{k=1}^{N} H(\nu,\xi_k)\ \Delta\xi_k\ \delta(t-\xi_k). \qquad (17)$$

In terms of this the system may be synthesized also in the configuration of figure 1 with $h_2(y,\xi_k)$'s replaced by $H(\nu,\xi_k)$'s.

Thus it is seen that the model of figure 1 may be used for synthesis when the system specification is given as either impulse response or H-system function.

(ii) $H(\nu, t)$ may also be approximated by

$$H^*(\nu, t) = \sum_{k=1}^{N} H(\beta_k, t)\, \Delta\beta_k\, \delta(\nu - \beta_k). \tag{18}$$

In terms of this, the approximate output is given by

$$v^*(t) = \sum_{k=1}^{N} A_k(t)\, v_k(t) \tag{19}$$

where $$v_k(t) = \int_{-\infty}^{\infty} H_k(\nu)\, U(\nu)\, e^{j2\pi\nu t}\, d\nu$$

$$H_k(\nu) = \Delta\beta_k\, \delta(\nu - \beta_k)$$

and $$A_k(t) = H(\beta_k, t)$$

Equation (19) indicates that the system may be synthesized in the form of model of figure 3. The time-varying gain amplifiers may require multipliers and function generators or amplitude modulators or sequentially gated summing amplifiers.

The time invariant subsystems may be approximated by bandpass filters with narrow passband and are independent of the desired system functions being approximated. Thus, the same time invariant subsystems may be used for modelling of systems which have the same range of input frequencies of interest, but may differ from one another in other respects.

3. Model from G-system function $G(\omega, \tau)$ where ω is the output frequency in cycles per second.

(i) The system function $G(\omega, \tau)$ may be approximated by

$$G^*(\omega, \tau) = \sum_{k=1}^{N} G(\omega, \xi_k)\, \Delta\xi_k\, \delta(\tau - \xi_k). \tag{20}$$

In terms of this approximation the approximate output spectrum is given by

$$V^*(\omega) = \sum_{k=1}^{N} G_k(\omega)\, x_k(\omega) \tag{21}$$

where $G_k(\omega) = \Delta\xi_k G(\omega, \xi_k)$

and $X_k(\omega) = x(\xi_k)\, e^{-j2\pi\omega\xi_k}$

which has the inverse Fourier transform

$$x_k(t) = x(\xi_k)\ \delta(t-\xi_k).$$

Thus, it is seen that the approximate output spectrum may be obtained by the configuration of figure 4.

This model consists of a sequential switch, time invariant subsystems and an adder and is dual to the one of figure 1.

In figure 4 the input time function to the k-th filter is an impulse $x(\xi_k)\ \delta(t-\xi_k)$. Such an impulse is approximated by means of the sequential switch. The switch connects at time ξ_k the k-th filter to the input which at that instant has the value $x(\xi_k)$.

The discussion given in connection with figure 1 is also applicable here

(ii) $G(\omega,\tau)$ may also be approximated by

$$G^*(\omega,\tau) = \sum_{k=1}^{N} G(\beta_k,\tau)\,\Delta\beta_k\ \delta(\omega-\beta_k). \tag{22}$$

In terms of this the approximate output spectrum is given by

$$V^*(\omega) = \sum_{k=1}^{N} X_k(\omega)\ Y_k(\omega) \tag{23}$$

where $X_k(\omega) = \Delta\beta_k\ \delta(\omega-\beta_k)$

and $Y_k(\omega) = \int_{-\infty}^{\infty} G(\beta_k,\tau)\ u(\tau)e^{-j2\pi\omega\tau}d\tau.$

The corresponding model is as shown in figure 5.

To realize this model time-varying gain amplifiers, fixed subsystems and an adder are required. Thus, the time-varying portion has the disadvantage of requiring multipliers and function generators, amplitude modulators or sequentially gated summing amplifiers. But the time invariant subsystems may be approximated by bandpass filters with narrow pass bands. They are independent of the system to be realized and depend only on the range of output frequencies of the system under consideration.

4. Model from Bifrequency System Function $\Gamma(\mu,\nu)$, where μ is is the frequency of system variation in cycles per second.

The bifrequency system function $\Gamma(\mu,\nu)$ may be approximated by

$$\Gamma^*(\mu,\nu) = \sum_{k=1}^{N} \Gamma(\mu,\beta_k)\ \Delta\beta_k\ \delta(\nu-\beta_k) \tag{23}$$

In terms of this, the approximate output spectrum is obtained as

$$V^*(\mu) = \sum_{k=1}^{N} \Gamma(\mu,\beta_k)\ \Delta\beta_k\ U(\beta_k). \tag{24}$$

The corresponding model is shown in figure 6 in which

$$\Gamma_k(\mu) = \Gamma(\mu,\beta_k)$$

and $\quad H_k(\nu) = \Delta\beta_k\ \delta(\nu-\beta_k)$.

This model is completely in the frequency domain. The subsystems $H_k(\nu)$ pass signals of a specific band of frequencies $\Delta\beta_k$. Subsystems $\Gamma_k(\mu)$'s represent the effect of the time variation of the system.

General discussion:

The development of these models does not require any time or frequency constraint on the system itself or the input or the output of the system.

Thus, these models are valid for systems with or without any such restrictions imposed on them. But the models are considerably simplified if some suitable restrictions are imposed on the system. For a particular restriction there is a particular model which proves more advantageous than others.

(1) In the case of input time restriction fewer terms are needed in the approximation formula (20) and in the corresponding

model shown in figure 4 also fewer subsystems are required.

(2) In the case of input frequency constraint the approximation formula (18) requires fewer terms and consequently the corresponding model of figure 3 also requires less parallel elements.

(3) In the case of output time constraint the model of figure 1 is simplified. An interesting case arises when $h_2(y, t)$ or $H(\nu, t)$ is periodic function of time. In such case the corresponding function need be approximated only over one period. Further, if the function is symmetric about the peak value then it is sufficient to approximate it for only half a period. For remainder of the time the same time invariant subsystems may be used by suitable programming of the switch. In some cases it may be even sufficient to approximate it for a quarter of a period.

(4) In the case of output frequency constraint the model shown in figure 5 is the advantageous one.

In the development of the system models by means of plane impulse train approximation, the aim has been to approximate a given characteristic function e.g., $h_2(y, t)$, $H(\nu, t)$ etc. by the corresponding approximate characteristic functions $h^*_2(y, t)$, $H^*(\nu, t)$ etc. respectively. This purpose has been achieved in the indirect way of approximating the actual output $v(t)$ of the system by the output $v^*(t)$ of the model for a general class of inputs $u(t)$.

Thus, considering the case of impulse response $h_2(y, t)$, for example, it is seen that $h_2(y, t)$ is approximated by $h^*_2(y, t)$ which consists of train of delayed plane impulses. So, the error involved in approximating $h_2(y, t)$ by $h^*_2(y, t)$ is obviously large. But for a general class of inputs the error involved in approximating the actual output $v(t)$ of the system by the output $v^*(t)$ of the model is reasonably small.

However, this method is generally good for relatively smooth inputs. In the case of impulse-like input the error is considerable, unless $h_2(y, t)$ also is impulse-like.

A question of importance is what should be the number of plane impulses required for any particular case. This number N will depend on the amount of allowable error at the output of the system. Considering the impulse response $h_2(y,t)$, for example, the approximate output is given by equation (12). Then a suitable error criterion may be chosen and N may be determined such, that the corresponding error satisfies the criterion. Thus, if it is required that the integral square error be less than certain preassigned positive number ϵ, then N should be chosen such that the integral

$$I = \int_{-\infty}^{\infty} \left| \int_{o}^{\infty} h_2(y,t) \; u(t-y)dy - \sum_{k=1}^{N} \Delta\xi_k v_k(t)\delta(t-\xi_k) \right|^2 dt \tag{25}$$

is less than ϵ .

The main advantage of the models developed from the method of plane impulse train approximation lies in their simplicity.

Example:

Consider the problem of synthesis of a system with H-system function

$$H(s,t) = \frac{1}{s+1+\rho\cos\omega_o t} \qquad o < \rho < 1 \tag{26}$$

where s is the complex input frequency.

The actual outputs of this system to unit step inputs delayed by different amounts are calculated.

The system is simulated on an analog computer by means of equation (17) and the corresponding model of figure 1. The outputs for unit step inputs delayed by different amounts of time are recorded.

The ideal (calculated) and experimental responses are shown in figures 7, 8 and 9 for comparison.

References:

1. L. A. Zadeh, Time-varying Networks, I., Proceedings of I. R. E. vol. 49, No 10, pp 1488-1503, Oct. 1961.

2. L. A. Zadeh, Frequency Analysis of Variable Networks, Proceedings of I. R. E., vol. 38, pp 291-299, March, 1950.

3. T. Kailath, Sampling Models For Linear Time-variant Filters. Technical Report No 352, May 25, 1959, M. I. T. Research Laboratory of Electronics, Cambridge, Massachussets, U. S. A.

4. Allen Gersho, Properties of Time-varying Linear Systems. M. Sc. Thesis, 1961, Cornell University Ithaca, N. Y. U. S. A.

5. L. A. Zadeh, and C. A. Desoer, Linear System Theory, McGraw-Hill Book Company, New York, 1963.

6. I. Matyash, Methods of Analog Computer Solution of Linear Differential Equations With Variable Coefficients, Automation and Remote Control, Vol. 20, No 7, pp 813, July, 1959.

7. L. Weiss, Time-varying Systems with Separable System Functions, The Johns Hopkins University Department of Electrical Engineering, Baltimore 18, Maryland, U. S. A.

8. S. V. Malchikov, On the Synthesis of Linear Automatic Control Systems with Variable Parameters, Automation and Remote Control, Vol. 20, No 12 December, 1959.

9. A. R. Stubberud, Synthesis of Linear Non-stationary Feedback Systems, I and II, IEEE Transactions on Applications and Industry, July, 1963, pp 186-196.

10. A. R. Stubberud, and E. R. Stear, Nonstationary Signal Flow Graph Theory, Report No 60-64, Department of Electrical Engineering, U. C. L. A. California, Nov. 1960.

11. J. B. Cruz, Jr. Synthesis of Linear Time-varying Systems, Ph. D. Thesis, University of Illinois, Urbana, U. S. A. 1959.

12. J.B. Cruz Jr. and M.E. Van Valkenburg, The Synthesis of Models for Time-varying Linear Systems, Proceedings of the Symposium on Active Networks and Feedback Systems, Vol. 10, Polytechnic Institute of Brooklyn, N.Y. 1960, pp 527-544.

13. P.A. Bello, Characterization of Randomly Time-variant Linear Channels. Advanced Communications Research Report No. 2, August 31, 1962.

14. J.B. Cruz Jr., A Generalization of the Impulse Train Approximation for Time-varying Linear System Synthesis in the Time Domain. IRE Transactions on Circuit Theory, Vol. CT-6 No 4, December, 1959.

15. G.S. Glinski and M. Arozullah, Synthesis of Linear Time-varying Systems, Technical Report No 64-11, June 1964, Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario Canada.

16. M. Arozullah, Synthesis of Linear Time-varying Systems, M.Sc. Thesis, 1964, Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario Canada.

17. J.G. Truxal, Numerical Analysis for Network Design, I.R.E. Transactions on Circuit Theory, CT-1, pp 49-60, September, 1954.

18. R.E. Kalman, Mathematical Description of Linear Dynamical Systems, Journal of Society of Industrial and Applied Mathematics. Ser. A, vol. 1, No 2, 1963, pp 153-192.

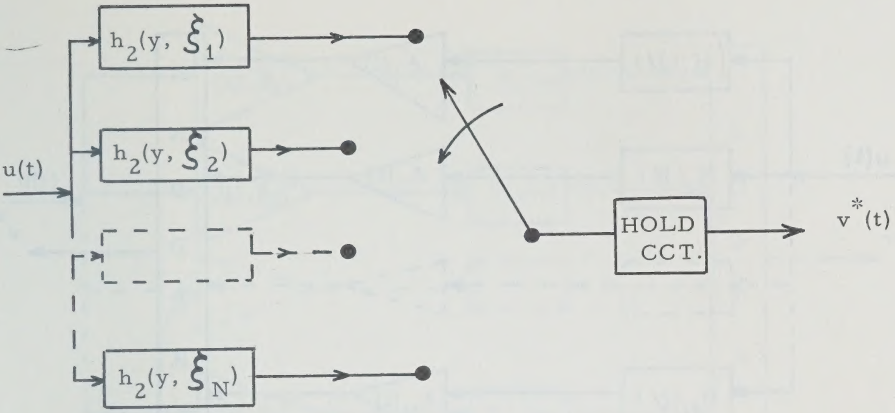

u(t)
$h_2(y, \xi_1)$
$h_2(y, \xi_2)$
$h_2(y, \xi_N)$
HOLD CCT.
$v^*(t)$

Figure 1. Model from $h_2(y, t)$.

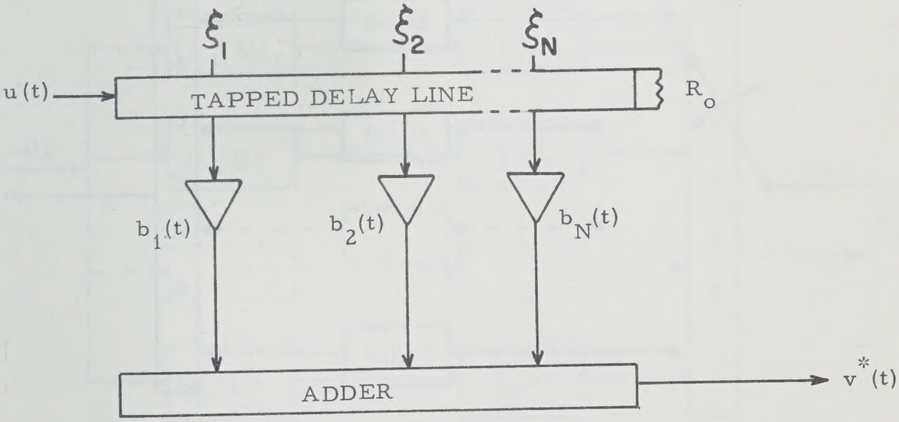

ξ_1
ξ_2
ξ_N
u(t)
TAPPED DELAY LINE
R_o
$b_1(t)$
$b_2(t)$
$b_N(t)$
ADDER
$v^*(t)$

Figure 2. Model from $h_2(y, t)$.

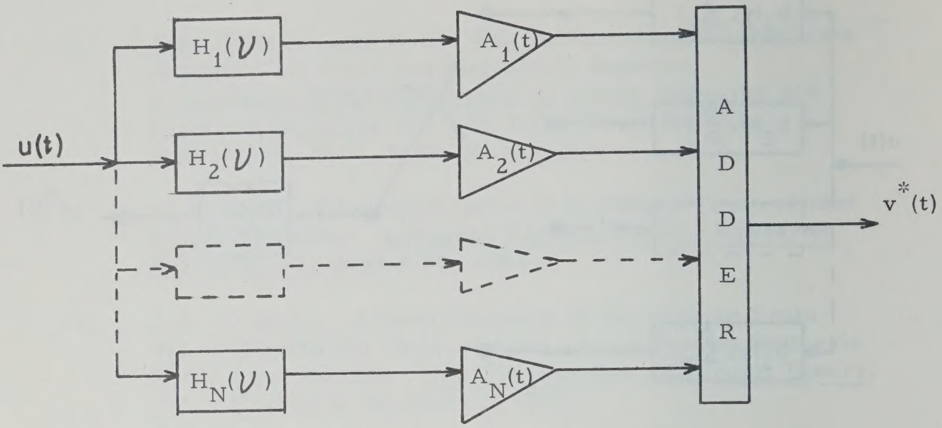

u(t)
$H_1(\nu)$
$H_2(\nu)$
$H_N(\nu)$
$A_1(t)$
$A_2(t)$
$A_N(t)$
A
D
D
E
R
$v^*(t)$

Figure 3. Model from $H(\nu, t)$.

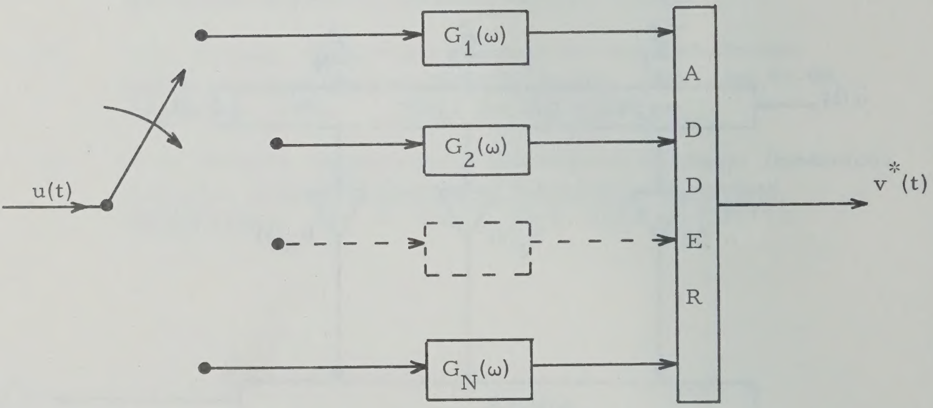

u(t)
$G_1(\omega)$
$G_2(\omega)$
$G_N(\omega)$
A
D
D
E
R
$v^*(t)$

Figure 4. Model from $G(\omega, \tau)$.

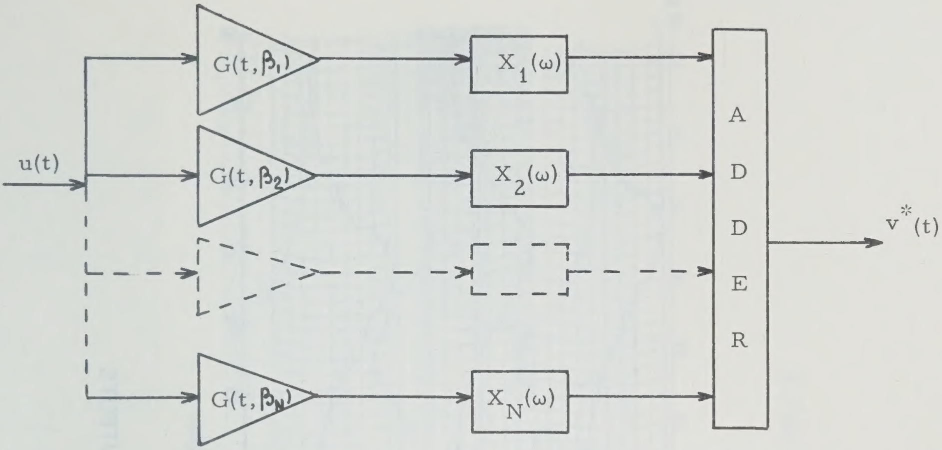

u(t)
G(t, β1)
G(t, β2)
G(t, βN)
X1(ω)
X2(ω)
XN(ω)
ADDER
v*(t)

Figure 5. Model from $G(\omega, \tau)$.

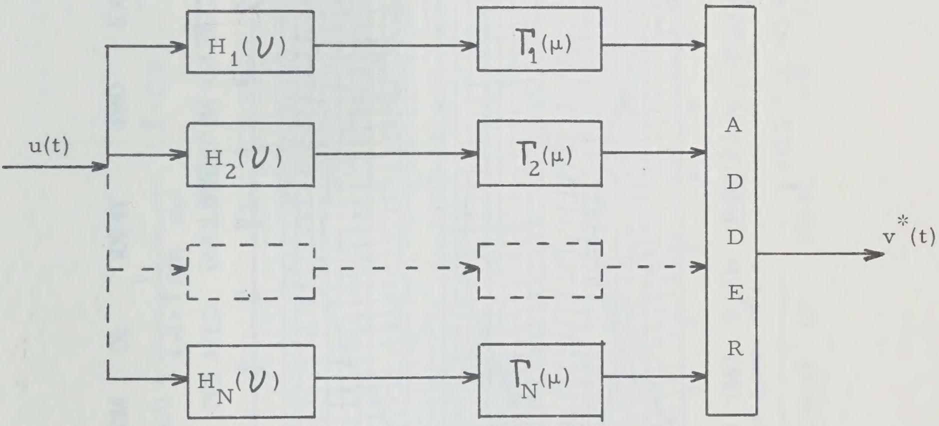

u(t)
H1(ν)
H2(ν)
HN(ν)
Γ1(μ)
Γ2(μ)
ΓN(μ)
ADDER
v*(t)

Figure 6. Model from $\Gamma(\mu, \nu)$.

$$H(s,t) = \frac{1}{s+1+\rho \cos \omega_0 t} \quad ; \ \rho = 0.5 \ ; \ \tau = 0 \text{ sec.}$$

INPUT : DELAYED UNIT STEP $U(t-\tau)$; $\omega_0 = 1.047$ radians /sec.

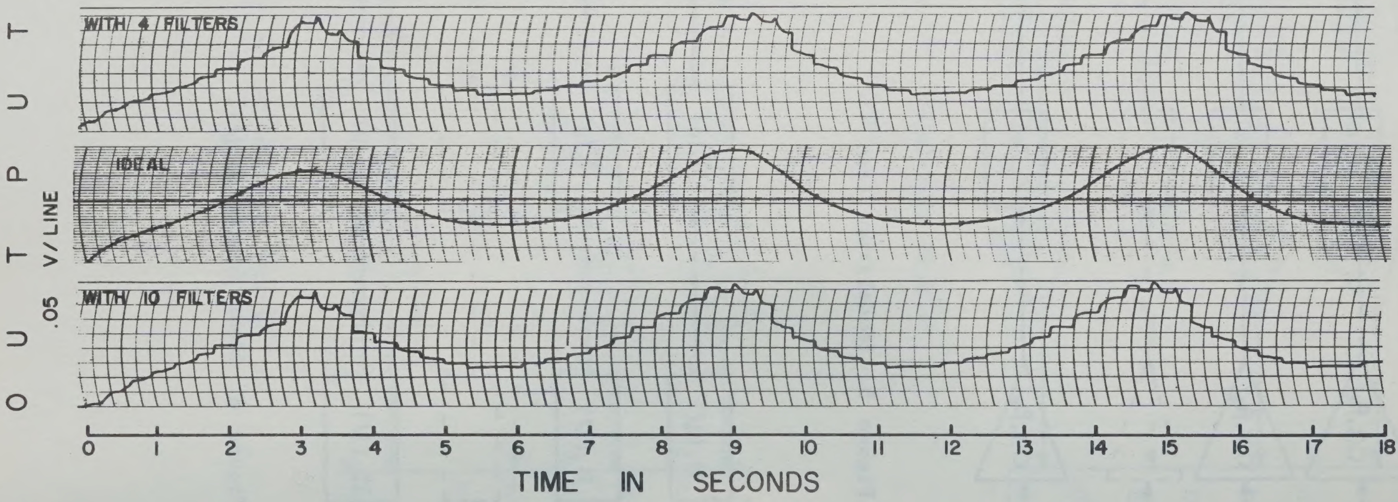


Fig. 7. COMPARISON OF IDEAL AND EXPERIMENTAL OUTPUTS

$$H(s,t) = \frac{1}{s+1+\rho\cos\omega_0 t} \quad ; \rho = 0.5 \quad ; \tau = 1.5 \text{ sec.}$$

INPUT : DELAYED UNIT STEP $U(t-\tau)$; $\omega_0 = 1.047$ radians/sec.

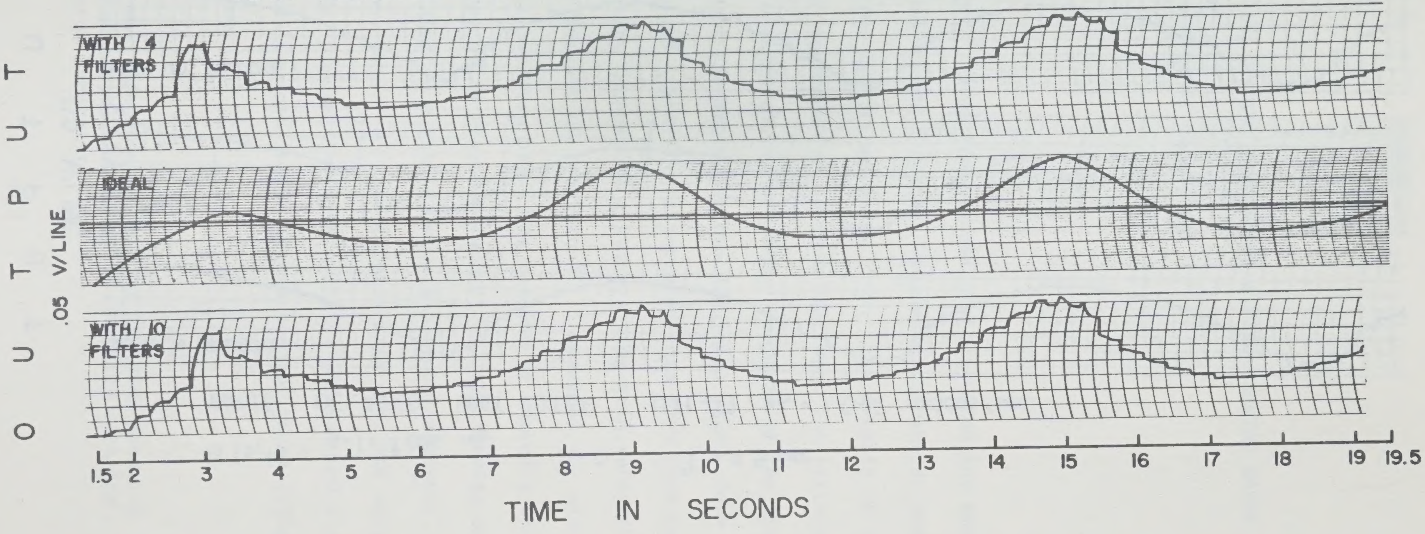


Fig. 8. COMPARISON OF IDEAL AND EXPERIMENTAL OUTPUTS

$$H(s,t) = \frac{1}{s+1+\delta\cos.\ \omega_0 t}\ ;\ \delta = 0.5\ ;\ \tau = 4.5\ \text{sec.}$$

INPUT : DELAYED UNIT STEP $U(t-\tau)$; $\omega_0 = 1.047$ radians/sec.

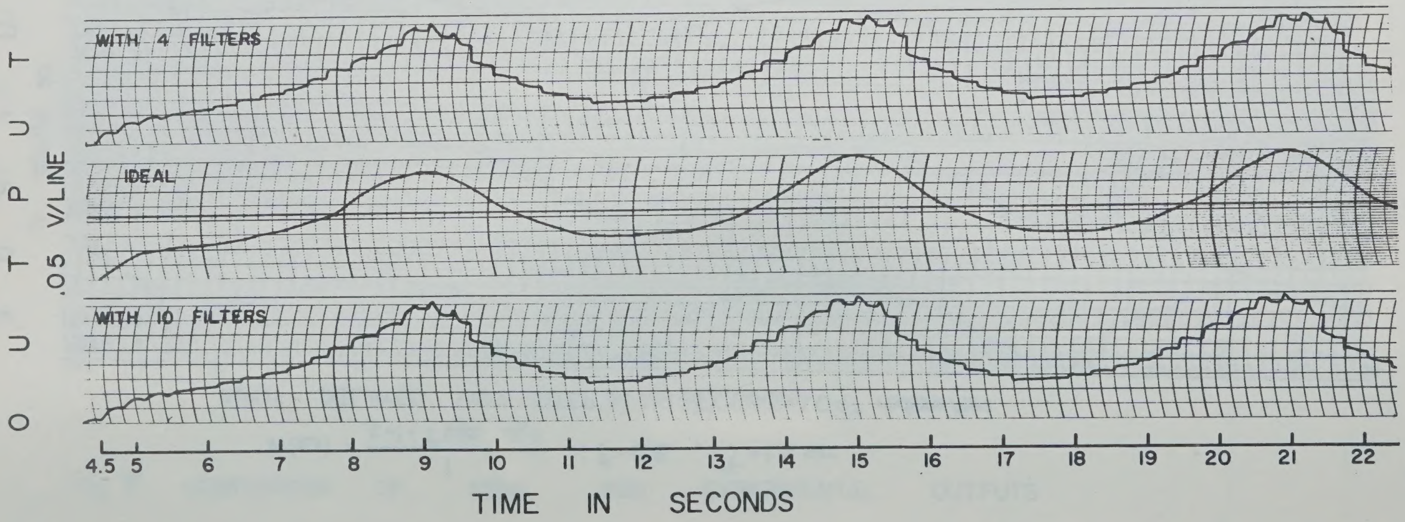


Fig. 9. COMPARISON OF IDEAL AND EXPERIMENTAL OUTPUTS.

ON THE STABLILITY OF SYSTEMS CONTAINING A TIME-VARYING GAIN*

R. W. Brockett and L. J. Forys
Massachusetts Institute of Technology
Cambridge, Massachusetts

INTRODUCTION

In 1961 Popov(1) developed a stability criterion for the class of nonlinear systems which contain a single instantaneous nonlinear gain in an otherwise linear, constant loop. The details of his theorem are not so important as is that fact that the criterion is stated entirely in terms of the frequency response of the linear part of the loop and the bounds on the nonlinear gain. Subsequently a number of stability criteria of this type have been developed for systems containing a time-varying gain. Particularly important are the results of Bongiorno(2-3) obtained for systems with a periodically varying gain, the results of Narendra and Goldwyn(4) obtained using Liapunov theory, and the results of Zames(5) and Sandberg(6) obtained using functional analysis. In this paper Liapunov theory is used to establish new, and more flexible, frequency domain stability criteria and the modified polar plot of Popov is used in the application of these criteria to specific systems.

---

* This research was supported in part by the Army Research Office under Contract No. DA-ARO(D)-31-124-G193 with the Electronic Systems Laboratory, M.I.T.

PRELIMINARIES AND NOTATION

Consider the feedback system containing a single time-varying gain shown in Figure 1. Such a model might represent a particular type of control system, a parametric amplifier, a part of a communications link, or a number of other physical devices. If the variable x is related to the response y by the equation

$$q(D)x = y; \quad D \equiv d/dt \tag{1}$$

then the differential equation describing the motion of x is simply

$$p(D)\, x + f(t)q(D)x = 0. \tag{2}$$

If all solutions of this equation are bounded then clearly y is also bounded; that is, it is possible to study the stability of the closed-loop system shown in Figure 1 by studying the stability of the differential Equation 2.

It will be assumed throughout that the transfer function in the forward part of the loop is of the form $G(s) = q(s)/p(s)$ with $q(s)$ and $p(s)$ being of the form

$$q(s) = q_{n-1}s^{n-1} + q_{n-2}s^{n-2} \ldots + q_1 s + q_0 \tag{3}$$

$$p(s) = s^n + p_{n-1}s^{n-1} + p_{n-2}s^{n-2} \ldots + p_1 s + p_0. \tag{4}$$

Any of the q_i and p_i may be zero, however the degree of $p(s)$ always exceeds that of $q(s)$ by at least one.

For some purposes it is convenient to have a representation of Equation 2 in the first order vector form

$$\dot{\underline{x}} = F(\underline{x}, t). \tag{5}$$

It is easily verified that if $\underline{x}$ is an n-vector whose i^{th} component is $D^{i-1}x$ then Equation 5 is equivalent to Equation 2 provided that the i^{th} component of $F(\underline{x},t)$ is x_{i+1} for $1 \leq i \leq n-1$

with the last component being given by

$$F_n(\underline{x},t) = -p_{n-1}x_n \cdots -p_0x_1 -f(t)(q_{n-1}x_n \cdots -q_0x_1). \qquad (6)$$

A first order vector equation such as 5 is said to be stable (in the sense of Liapunov) if one can insure that the solution $\emptyset(t,\underline{x}_0,t_0)$ corresponding to an initial value $\underline{x}_0$ at time t_0, remains close to origin for all time simply by requiring $\underline{x}_0$ to be sufficiently small. (For a precise definition see Hahn,(7) LaSalle and Lefschetz,(8) or Kalman and Bertram.(9)) Ordinarily, stability in this sense implies nothing about the behavior of the solutions for large initial conditions; but since the equations to be considered are linear, stability in this sense implies boundedness for any initial conditions.

The basic tool which will be used to establish stability is the following theorem of Liapunov.

Theorem 1: The time-varying system (5) is stable if there exists a continuous positive definite function $V(\underline{x})$, having continuous first partial derivatives with respect to $\underline{x}$, such that the derivative of $V(\underline{x})$ is nonpositive along any solution of 5.

A proof of this can be found in Reference 8. Notice that for the special case when Equation 5 is generated from an n^{th} order scalar equation in the way described above, this theorem says that stability can be established by finding a positive definite function of x and its first n-1 derivatives whose derivative is nonpositive along solutions of Equation 2.

The Liapunov functions to be used here consist of a time invariant part plus time varying terms. The following theorem, which is a slight extension of a theorem appearing in Reference

10, will be used in generating the time invariant part.

Theorem 2: Let $\underline{x}$ be a state vector for Equation 2 whose components are x and its first n-1 derivatives, and let Ev and $(\ )^{-}$ denote "even part" and "right half-plane spectral factor", respectively. If q(s)/p(s) is positive real, and if q(s), p(s), and $(Evq(s)p(-s))^{-}$ have no common zeros in the right half-plane or on the imaginary axis, then the line integral

$$V(\underline{x}) = \int_{t(0)}^{t(\underline{x})} q(D)xp(D)x - ((Evq(D)p(-D))^{-}x)^{2}dt \tag{7}$$

is independent of path and is a positive definite function of $\underline{x}$.

Proof: In Reference 10 it is shown that the given line integral is independent of path. To show that it is positive definite it is convenient to appeal to a theorem of Weinberg and Slepian(11) which says that if q(s)/p(s) is positive real and if q(s) and p(s) do not have common right half-plane or imaginary axis zeros then the zeros of q(s) + p(s) lie in the left half-plane. Let $\phi(\underline{x})$ denote a solution of

$$p(D)x + q(D)x = 0 \tag{8}$$

which starts at the state x at t = 0. Evaluated along this particular path $V(\underline{x})$ is given by

$$V(\underline{x}) = \int_{0}^{\infty} (p(D)\phi)^{2} + (Evq(D)p(-D))^{-}\phi)^{2}dt. \tag{9}$$

This shows that $V(\underline{x})$ is at least positive semi-definite. If $V(\underline{x})$ were to be zero for some nonzero $\underline{x}$ then $p(D)\phi$ and $(Evq(D)p(-D))^{-}\phi$ would have to vanish identically along some solution of Equation 8. Clearly this could only happen if p(s), q(s), and $(Evq(s)p(-s))^{-}$ have a common factor -- contrary to our assumptions.

FREQUENCY DOMAIN STABILITY CRITERIA

Physical considerations suggest that if a passive network is terminated with a time-varying resistance $f(t)$ which is positive for all values of t then the system should be stable. (See Figure 2.) Since the equation of motion for a system consisting of an impedance $p(s)/q(s)$ terminated with a resistance $f(t)$ is identical to Equation 2 it follows that Equation 2 should be stable if $f(t)$ is nonnegative and $q(s)/p(s)$ is positive real. By exploring the effect of adding and subtracting a fixed shunt resistance of value k one is led to conjecture the following theorem.

Theorem 3: The system in Figure 1 is stable if $G(s) + 1/k$ is a positive real function whose real part is not identically zero and if $0 \leq f(t) \leq k$.

Results equivalent to this have been derived by Bongiorno, Narendra and Goldwyn, Zames, and Sandberg in the papers cited in the references. Several alternative forms are also given by Narendra and Goldwyn and the relationship between this statement of the result and that due to Bongiorno(2) is also discussed in their report. Using theorems 1 and 2 it is possible to give a short proof.

Proof: Assume that $G(s) + 1/k$ and $f(t)$ have the properties described in the theorem. Multiply Equation 2 by $q(D)\,x + p(D)x/k$ and subtract $((Ev(q(D) + p(D)/k)p(-D))^{-}x)^2$ from each side to get

$$\begin{aligned} q(D)\,x\,p(D)x + p(D)xp(D)x/k - ((Ev(q(D) + p(D)/k)p(-D))^{-}x)^2 &= \qquad (10) \\ -((Ev(q(D) + p(D)/k)p(-D))^{-}x)^2 - f(t)(q(D)x)^2 &- f(t)q(D)xp(D)x/k. \end{aligned}$$

Using the fact that $p(D)x = -f(t)\,q(D)x$, it follows that an integration of Equation 10 from $t(0)$ to $t(\underline{x})$ gives

$$\int_{t(0)}^{t(\underline{x})} q(d)x\ p(D)xp(D)x/k - ((Ev(q(D) + p(D)/k)p(-D)^{-}x)^{2}dt = \qquad (11)$$

$$\int_{t(0)}^{t(\underline{x})} -((Ev(q(D) + p(D)/k)p(-D))^{-}x)^{2} - f(t)(1-f(t)/k)(q(D)x)^{2}dt\ .$$

Now from theorem 2 the quantity on the left is a positive definite function of $\underline{x}$, its derivative is clearly given by the integrand on the right. Since $0 \leq f(t) \leq k$ it follows that the integrand is nonpositive and hence theorem 1 implies stability. Q.E.D.

Applications of this theorem to specific systems will be considered in a later section, however it is immediately clear that it makes rather severe restrictions on the linear part of the system compared with what should be predicted if the gain were fixed. That is, the Nyquist criterion predicts stability for fixed gain if the Nyquist plot avoids the part of the negative real axis lying to the left of the $-1/k$ point whereas theorem 3 predicts stability only if the Nyquist plot avoids the half-plane lying to the left of the vertical line passing through the $-1/k$ point. The following generalization of theorem 3 partially circumvents this difficulty.

Theorem 4: The system in Figure 1 is stable if there exists α , β , and k such that $(G(s) + 1/k)\ (1 + \alpha s)/(1 + \beta s)$ is positive real, $0 \leq f(t) < k$, and

$$\dot{f}(t)/f(t) \leq 2(1-f(t)/k)\min(1/\alpha,\ 1/\beta). \qquad (12)$$

Before proving the theorem it is necessary to establish the following lemma.

Lemma: If $n(D)x = 0$ is asymptotically stable then the stability of the equation

$$p(D)n(D)x + f(t)q(D)n(D)x = 0 \qquad (13)$$

implies the stability of Equation 2.

Proof of Lemma: If $\phi(t)$ is a solution of Equation 2 then $n(D)\phi(t)$ is a solution of Equation 13. If $\phi(t)$ is unbounded and if $n(s)$ has all its zeros in the right half-plane then $n(D)\phi(t)$ is unbounded also. (This is a consequence of the fact that for asymptotically stable systems bounded inputs give rise to bounded outputs.) Consequently if all the solutions of Equation 13 are bounded then those of Equation 2 must be also.

Now return to the proof of the theorem.

Proof: In view of the lemma it is enough to prove that the given assumptions imply the stability of the equation

$$(1 + \beta D)p(D)\, x + f(t)\,(1 + \beta D)q(D)x = 0. \tag{14}$$

First consider the case where $k = \infty$. Multiply Equation 14 by $(1 + \alpha D)q(D)x$ and subtract $((\mathrm{Ev}(1 + \alpha D)q(D)\,(1-\beta D)p(-D))^{-}x)^2$ from each side to get

$$\begin{aligned}&(1 + \alpha D)q(D)x(1 + \beta D)p(D)x - ((\mathrm{Ev}(1+\alpha D)q(D)(1-\beta D)p(-D))^{-}x)^2 = \\ &-f(t)(1+\alpha D)q(D)x(1+\beta D)q(D)x - ((\mathrm{Ev}(1+\alpha D)q(D)(1-\beta D)p(-D)^{-}x)^2.\end{aligned} \tag{15}$$

If α exceeds β then expand the first term on the right side as

$$\begin{aligned}&f(t)\,(1 + \alpha D)q(D)x\,(1 + \beta D)q(D)x = \\ &f(t)\,((1 + \alpha D)q(D)x)^2\,\beta/\alpha + f(t)(1 - \beta/\alpha)q(D)x(1 + \alpha D)q(D)x.\end{aligned} \tag{16}$$

By using Equation 16 and the integration-by-parts identify

$$\begin{aligned}&\int_{t(0)}^{t(\underline{x})} f(t)q(D)x\,(1 + \alpha D)q(D)x\,dt = \\ &1/2\,\alpha\, f(t)(q(D)x)^2 + \int_{t(0)}^{t(\underline{x})} (f(t) - 1/2\,\alpha\,\dot{f}(t))(q(D)x)^2 dt\end{aligned} \tag{17}$$

It follows that

$$\begin{aligned}&\int_{t(0)}^{t(\underline{x})} \Big[(1 + \alpha D)q(D)x\,(1 + \beta D)p(D)x - \\ &((\mathrm{Ev}(1 + \alpha D)q(D)(1 - \beta D)p(-D))^{-}x)^2\Big]dt + 1/2\, f(t)(\alpha-\beta)(q(D)x)^2 =\end{aligned} \tag{18}$$

$$\int_{t(0)}^{t(\underline{x})} - ((EV(1 + \alpha D)q(D)(1 - \beta D)p(-D))^{-}x)^2 - f(t)((1 + \alpha D)q(D)x)^2\, \beta/\alpha - (f(t) - \alpha/2\, \dot{f}(t))(q(D)x)^2(\alpha - \beta)/\alpha \; dt$$

In view of theorem 2 and the assumptions which have been made of f(t) it follows that the quantity appearing on the left side of this equality is a positive definite function of $\underline{x}$. Its derivative, the integrand appearing on the right, is clearly nonpositive and thus by theorem 1 it follows that Equation 14 is stable. It is clear that α and β enter into Equation 15 in a symmetric way so that if β exceeds α an equivalent argument holds. Thus the special case where $k = \infty$ has been disposed of.

The case where f(t) is bounded by k can be treated by a simple transformation of the governing equations. If $0 \le f(t) < k$ then add and subtract $f(t)p(D)x/k$ from the left side of Equation 2 to get

$$(1 - f(t)/k)p(D)x + f(t)(q(D) + p(D)/k)x = 0. \tag{19}$$

Clearly this equation has the same solutions as Equation 2. Dividing through by $(1 - f(t)/k)$ gives

$$p(D)x + ((kf(t)/(k - f(t))(q(D) + P(D)/k)x = 0. \tag{20}$$

This equation also has the same solutions as Equation 2 and hence the same stability properties. From the special case proven above it follows that it is stable if there exists α and β such that $(1 + \alpha s)G(s)/(1 + \beta s)$ is positive real and if

$$\frac{d}{dt}\left(\frac{kf(t)}{k - f(t)}\right) \div \frac{kf(t)}{k - f(t)} \le 2 \min(1/\alpha, 1/\beta). \tag{21}$$

The indicated derivative has the value $k^2 \dot{f}(t)/(k - f(t))^2$; a simplification shows that inequality (21) is equivalent to inequality (12) and hence the theorem is proven.

Notice that this is a true generalization of theorem 3 in that theorem 3 corresponds to the special case where α and β are zero. It should be pointed out that Narendra and Goldwyn(4) establish the special case obtained by setting $\beta = 0$. Theorem 4 is not, however, the best theorem of type. Its hypothesis can be weakened slightly as the following theorem indicates.

Theorem 5: The system in Figure 1 is stable if there exists an RC or RL type driving point impedance Z(s) having poles at $-\beta_i$ and zeros at $-\alpha_i$ such that Z(s)(G(s) + 1/k) is positive real, $0 \leq f(t) < k$, and

$$\dot{f}(t)/f(t) \leq 2(1 - f(t)/k) \min (1/\alpha_i, 1/\beta_i). \qquad (22)$$

Since the proof of this is rather involved it will not be given here; a complete proof can be found in Reference 12. The basic idea of the proof is to let $n(D) = \prod_i (i + \beta_i D)$ and to study the equation

$$n(D)p(D)x + f(t)n(D)q(D)x = 0$$

in much the same way as Equation 14 was studied. The Liapunov function used consists of a quadratic part plus one time varying term for each pole of Z(s).

THE MODIFIED POLAR PLOT

Given a transfer function G(s), it is possible to find the largest value k such that G(s) + 1/k is positive real simply by examining the Nyquist polot of G(s); the minimum of the real part of $G(j\omega)$ is the most appropriate value for 1/k. The application of theorems 4 and 5 is not so straight forward however, because it is not clear from the Nyquist plot whether or not a transfer function can be made positive real by the multiplication of a factor $(1 + \alpha s)/(1 + \beta s)$. The modified polar plot used by Popov(1) is quite useful for this purpose.

Instead of plotting Im $G(j\omega)$ vs. Re $G(j\omega)$ as is done in the Nyquist plot, Popov suggested plotting ωIm $G(j\omega)$ vs. Re $G(j\omega)$. Having such a plot, it is a simple matter to determine if

$$\mathrm{Re}\,(1 + \alpha j\omega)G(j\omega) = \mathrm{Re}\,G(j\omega) - \alpha\,\omega\,\mathrm{Im}\,G(j\omega) \geq 0\,. \qquad (23)$$

Clearly a necessary and sufficient condition for this inequality to hold is that the modified polar plot should lie entirely to the right of a line passing through the origin with slope $1/\alpha$. Since testing the real part for non-negativity is the only really difficult thing involved in testing for positive realness, the modified polar plot gives almost all the information needed to apply theorem 4, including the minimum permissible value for α.

Of course the modified polar plot of $G(j\omega)$ also answers the same questions for the function $(1/(1 + \beta s))G(j\omega)$. In this case the real part will be nonnegative only if

$$\mathrm{Re}(1 = \beta j\omega)G(j\omega) = \mathrm{Re}\,G(j\omega) + \beta\,\omega\,\mathrm{Im}\,G(j\omega) \geq 0. \qquad (24)$$

Hence, the modified polar plot should lie to the right of a line passing through the origin, having slope $-1/\beta$.

In a similar way, $(1 + \alpha s)(G(s) + 1/k)$ is positive real for some α if the modified polar plot likes to the right of a line passing through the $-1/k$ point, having the slope $1/\alpha$. For the $(1/(1 + \beta s))(G(s) + 1/k)$ case the slope of the line must be $-1/\beta$.

The modified polar plots of some second, third, and fourth order systems will be given in the next section. The modified polar plot is similar to the Nyquist plot in that it crosses the real axis at the same points and at the same frequencies; the main differences are that it does not necessarily start at, or end up at, a multiple of 90° and that if one integration is present the locus does not go to infinity at low frequencies.

EXAMPLES

In this section the preceding theorems are applied to some specific systems. The results obtained using these methods are found to compare favorably with some obtained by others using *ad hoc* methods.

Example 1: Consider the second order system $G(s) = 1/(s^2 + as)$ and to which feedback through a gain $f(t)$ has been applied. The resulting equation of motion is

$$\ddot{x} + a\dot{x} + f(t)x = 0 \ . \tag{25}$$

For such a system, theorem 3 requires that $f(t)$ be bounded by a^2 since $1/a^2$ is the least $1/k$ can be if $G(s) + 1/k$ is to be positive real. On the other hand, if restrictions are placed on $\dot{f}(t)$ then the constraint on $f(t)$ can be relaxed. Consider the modified polar plot shown in Figure 3. Since the modified polar plot enters the origin with a slope $1/a$, and lies to the right of this line for all ω, it follows from theorem 4 that Equation 25 will be stable if $f(t)$ is positive and $\dot{f}(t)$ is less than $2\,a\,f(t)$. Of course there is a trade-off between these two extreme cases. This is shown graphically in Figure 3. The line m-o shown there defines a gain k and a slope α such that if $f(t) < k$ and $\dot{f}(t)/f(t) \leq (1 - f(t)/k)2/\alpha$ then the system will be stable. This can be expressed analytically by saying that the system will be stable if there exists k such that $0 \leq f(t) < k$ and $\dot{f}(t)/f(t) \leq (1 - f(t)/k)\cdot(k - a^2)/ak$.

Example 2: The corresponding third order system having no zeros and a pole at the origin has a transfer function $G(s) = 1/(s^3 + as^2 + bs)$ and an equation of motion of the form

$$\dddot{x} + a\ddot{x} + b\dot{x} + f(t)x = 0 \; . \qquad (26)$$

In this case the minimum value of the real part of $G(j\omega)$ is either $-4/(4ab - a^3)$ or $-a/b^2$ whichever is smaller, and therefore theorem 3 predicts stability only if

$$0 < f(t) < \max(ab - a^3/4, \; b^2/a). \qquad (27)$$

An examination of the modified polar plot shown in Figure 4 reveals that it passes through the $-1/ab$ point with a slope b/a and hence theorem 4 predicts stability if

$$\dot{f}(t)/f(t) \leq (1 - f(t)/ab) \; 2b/a \; . \qquad (28)$$

Again there is a trade-off between the constraint on $\dot{f}(t)$ and the constraint on $f(t)$ as shown by the line o-m in Figure 4.

In this case the result predicted by theorem 3 duplicates a result of Starzinskii quoted by Hahn (see Reference 7, page 40), and the result predicted by theorem 4 seems to be new.

Example 3: For the two examples treated above β was taken to be zero. As an example in which this is not the case, consider a fourth order system without zeros which has a pole at the origin. Let $G(s) = 1/(s^4 + as^3 + bs^2 + cs)$ so that the equation of motion becomes

$$\ddddot{x} + a\dddot{x} + b\ddot{x} + c\dot{x} + f(t)x = 0; \; ab - c > 0. \qquad (29)$$

The modified polar plot for this system is shown in Figure 5. As is indicated there, it passes through the point $-a^2/(abc - c^2)$ with a slope $ac/(ab - 2c)$, a number which may be negative. It lies to the right of the straight line passing through this point with the same slope if and only if $a^3 - (ab - c) \geq 0$. Thus by using theorem 4 it is possible to prove the stability of Equation 27 provided

$$\dot{f}(t)/f(t) \leq (1 - f(t)/k)\ 2\ ac/|ab - 2c| \ ;$$
$$k = (abc - c^2)/a^2 \tag{30}$$

and $a^3 - (ab - c) > 0$. Of course if $a^3 - (ab - c)$ is negative it will still be possible to use theorem 4 to prove stability but in this case it will be necessary to make further restrictions on f(t).

In the special case where the realpart of $G(j\omega)$ takes on its minimum value when the imaginary part of $G(j\omega)$ vanishes, the conditions given by theorem 3 are both necessary and sufficient. In general, however, the conditions given by theorems 3, 4, and 5 have only been proven to be sufficient so that if their conditions are violated there is no assurance that it will be possible to find an f(t) which drives the system unstable. Still, it is possible to get some idea about the types of systems which might be made unstable by time-varying gains by a closer examination of the classes of systems which these theorems can not treat.

Example 4: Consider a system with a double zero at the origin, say $G(s) = s^2/(s^3 + 2s^2 + 2s + 1)$. The Nyquist plot of this system avoids the negative real axis and thus the system is stable for all constant negative feedback, regardless of its magnitude. Yet, it will never be possible to find α and β such that $(1 + \alpha s)G(s)/(1 + \beta s)$ is positive real because the lowest power present in the numerator is s^2. This means that some bound must be applied to f(t) to insure stability and this bound decreases by permitting a larger f(t). By varying f(t) between 0 and 1 in an on-off fashion it should be possible to produce instability.

Another limitation associated with theorem 4 is that if the phase exceeds 90° for some frequency and if it falls below -90° for some other frequency, then it will be impossible to establish stability without bounding f(t). The reason is that $(1 + \alpha s)/(1+\beta s)$ either has a positive phase angle for all frequencies or else it has a negative phase angle for all frequencies. Thus if the phase of G(s) both exceeds 90° and falls below -90° it will not be possible for the phase angle of $(1 + \alpha s)G(s)/(1 + \beta s)$ to lie between 90° and -90° and hence it cannot be positive real.

A question of great interest is that of determining if the condition given by theorem 3 is a necessary one for the stability of Equation 2 for all f(t) between 0 and k. An answer to this might be obtained by a detailed analysis of systems with piecewise constant gains since this class of systems seems to be both quite general and at the same time capable of being analyzed exactly. The question of the necessity of theorem 4 is also of interest but this would seem to be a much harder problem.

CONCLUSIONS

The objective of this paper has been to develop frequency domain stability criteria for an important class of time-varying linear systems and to apply these criteria to specific equations. The principal new general results are given by theorems 4 and 5, and new stability results for second, third and fourth order equations have been given as examples. Although a number of fairly general examples have been given, these by no means exhaust the possibilities.

It has been shown that the modified polar plot of Popov is quite useful when applying theorems 3 and 4, however no equivalent tool seems to be avilable to aid in the application of the more general theorem 5. Another important area for future investigation is the generation of necessary conditions for stability of the type of systems under consideration here.

ACKNOWLEDGMENT

The authors are indebted to Professor G. Zames for many helpful suggestions. They would also like to acknowledge fellowship support from the Ford Foundation (R.W.B.) and the National Science Foundation (L.J.F.).

REFERENCES

1. V. M. Popov, "Absolute Stability of Nonlinear Systems of Automatic Control", Automation and Remote Control, Vol. 22, No. 8 (March, 1962) pp. 857-875. (Russian Original published in August, 1961.)

2. J. J. Bongiorno, Jr., "An Extension of the Nyquist-Barkhausen Stability Criterion to Linear Lumped-Parameter Systems with Time-Varying Elements", IEEE Trans. on Automatic Control (Correspondence) Vol. AC-8 (April, 1963) pp. 166-172.

3. J. J. Bongiorno, Jr., "Real-Frequency Stability Criteria for Linear Time-Varying Systems", Proceedings of the IEEE, Vol. 52, No. 7 (July, 1964) pp. 832-841.

4. K. S. Narendra and R. M. Goldwyn, "Generation of Quadratic Type Liapunov Functions for Linear Time-Varying Systems", Technical Report No. 421, Cruft Laboratory, Harvard University; November, 1963.

5. G. Zames, "Functional Analysis Applied to Nonlinear Feedback Systems", IEEE Trans. on Circuit Theory, Vol. CT-10, No. 3 (September, 1963) pp. 392-404.

6. I. W. Sandberg, "A Frequency-Domain Condition for the Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element", Bell Systems Technical Journal, (to appear).

7. Wolfgang Hahn, Theory and Applications of Liapunov's Direct Method. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1963.

8. J. LaSalle and S. Lefschetz, Stability by Liapunov's Direct Method. New York: Academic Press, 1961.

9. R. E. Kalman and J. E. Bertram, "Control Systems Analysis and Design via the 'Second Method' of Liapunov", J. of Basic Engr., Trans. ASME, Vol. 82 (June, 1960) pp. 371-393.

10. R. W. Brockett, "On the Stability of Nonlinear Feedback Systems", Preprints, 1964 Joint Automatic Control Conference (June 1964) pp. 288-294.

11. L. Weinberg and P. Slepian, "Positive Real Matrices", Journal for Mathematics and Mechanics, Vol. 9, No. 1 (1960) pp. 71-83.

12. L. J. Forys, Stability of Time-Varying Systems by Liapunov's Method, S.M. Thesis, Dept. of Electrical Engr., Mass. Institute of Technology. (to appear)

13. Blath, Van der Pol and M.J.O. Strutt, "On the Stability of Mathieu's Equation", Phil. Mag., Vol. 18, No. 5 (May, 1928) pp. 18-38.

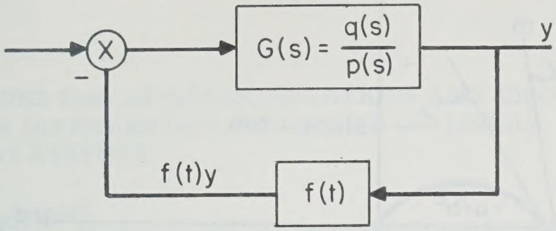

X
–
G(s) = q(s)/p(s)
y
f(t)y
f(t)

Figure 1: The Block Diagram of the System under Consideration

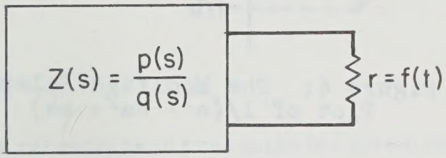


Figure 2: The Network Equivalent of Figure 1

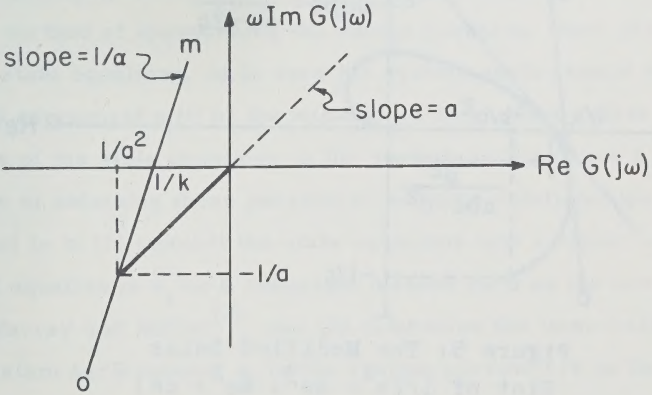

ωIm G(jω)
slope = 1/a
m
slope = a
1/a²
Re G(jω)
1/k
–1/a
0

Figure 3: The Modified Polar Plot of $1/(s^2 + as)$

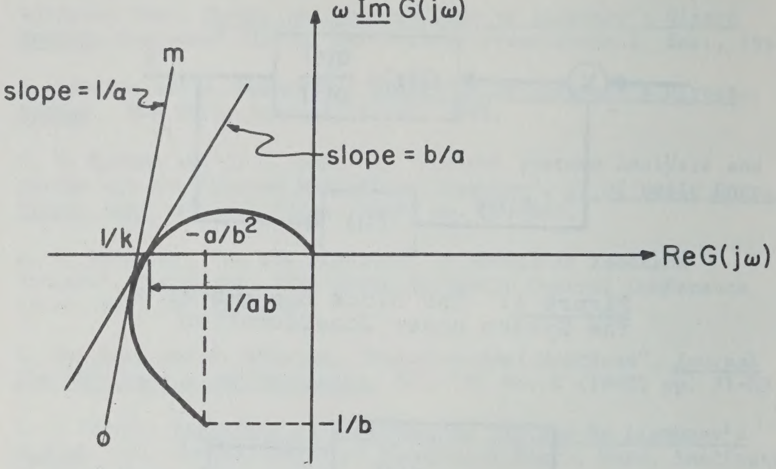

ω Im G(jω)
m
slope = 1/a
slope = b/a
1/k
−a/b²
Re G(jω)
1/ab
o
−1/b

Figure 4: The Modified Polar Plot of $1/(s^3 + as^2 + bs)$

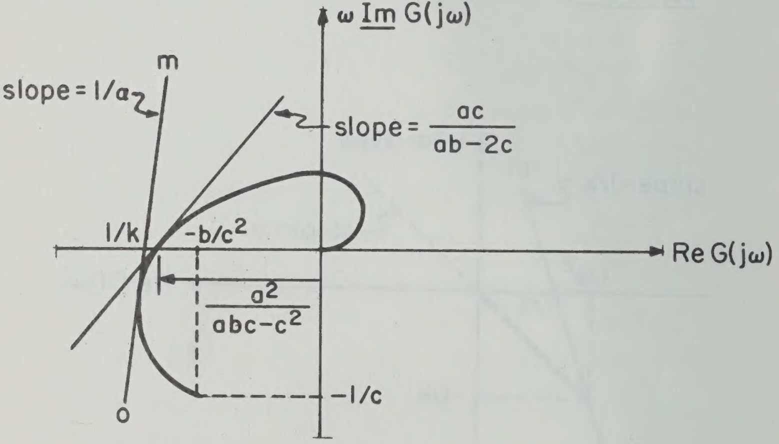

ω Im G(jω)
m
slope = 1/a
slope = ac/(ab−2c)
1/k
−b/c²
Re G(jω)
a²/(abc−c²)
o
−1/c

Figure 5: The Modified Polar Plot of $1/(s^4 + as^3 + bs^2 + cs)$

NEW PROPERTIES OF STATE EQUATIONS AND ASSOCIATED HIGHER ORDER DIFFERENTIAL EQUATIONS OF LINEAR, TIME-INVARIANT SYSTEMS

Robert H. Rogers
Bendix Systems Division, The Bendix Corporation, Ann Arbor, Michigan

David P. Brown
Department of Electrical Engineering, University of Wisconsin, Madison, Wisconsin

I. INTRODUCTION

Procedures have recently been presented for formulating the mathematical model of a physical system into the form $\dot{X} = AX + Q(t)$ where $X' = [x_1 x_2 \ldots x_n]$, $A = [a_{ij}]$, $Q'(t) = [q_1(t) q_2(t) \ldots q_n(t)]$ and the prime denotes the transpose. This set of differential equations is commonly referred to as the state equations.$^{(1)}$

One method of approaching the design problem, when presented with the state equations, is to view the system performance as one specified component $x_i(t)$ of the solution vector. Then relate the parameters of the state equations to the system performance $x_i(t)$. One technique of obtaining these parameter-solution relationships (design equations) is to (1) convert the state equations into a higher-order differential equation in x_i by a reduction method such as the one proposed by Murray and Miller$^{(2)}$, and (2) determine the inter-relationship of the system performance x_i to the system parameters in the higher-order differential equation. An obstacle to obtaining the desired parameter-solution relationships by this technique is that the reduction method may convert the state equations into a differential equation of r-order, where the number r is less than the number of state equations, i.e., $r \leq n$.$^{(3)}$

Coddington and Levinson(5) as well as others(6) have applied the transformation $[x_1 x_2 \ldots x_n]' = [yy^{(1)} \ldots y^{(n-1)}]'$ where $y^{(i)} = \frac{d^i}{dt^i} y$, to the state equations, where matrix A has the special form of a companion matrix and $Q'(t) = [O\ O \ldots O\ q_n(t)]$ to obtain an n-order differential equation. The n-order equation thereby obtained is referred to as "associated" with the system. A generalization of this concept of the associated differential equation is required in this paper to develop the general mathematical properties which link the solution of the state equations to the solution of the r-order differential equation obtained from the state equations. This concept results in two techniques for relating the parameters in the state equations (entries of matrix A and vector Q(t)) to one specified component $x_i(t)$ of the solution vector $\mathbf{X}(t)$. The mathematical properties developed, in particular the parameter-solution relationships, afford new tools for the design of linear systems.

II. HOMOGENOUS STATE EQUATIONS AND ASSOCIATED HIGHER ORDER DIFFERENTIAL EQUATIONS

It is proven below that higher order differential equations are obtained from state equations by a non-singular transformation, $\mathbf{X} = C^{-1} Y$ and for this case, the solution of the state equations and the solution of the higher order differential equations are linked by two well-defined properties. These properties are given in the following definition.

Definition 1: Consider the homogenous system of differential equations:

$$\dot{\mathbf{X}} = A\mathbf{X} \tag{1}$$

where $A = [a_{ij}]$ and a_{ij} is a constant, $\mathbf{X}' = [x_1, x_2 \ldots x_n]$. An r-order homogenous differential equation

$$\frac{d^r y}{dt^r} = \sum_{i=1}^{r} a_i \frac{d^{r-i} y}{dt^{r-i}} \tag{2}$$

is associated with (1) if,

(a) for F_m a fundamental matrix of (1), there exists non-singular matrices C and D such that the (i, i), (i, i+1), ..., (i, i+r-1) entries in CF_mD, for some i, i = 1, 2, ...n, are a fundamental set of (2) and

(b) for $y_1, y_2, \ldots, y_r$ a fundamental set of (2), the vectors $C^{-1}Y_j$ where $Y_j' = [0\;0 \ldots y_j y_j^{(1)} \ldots y_j^{(r-1)}\;0\;0 \ldots 0]$, j = 1, 2, ..., r, are r linearly independent solutions of (1) on the open interval I defined by I = (-∞, ∞).

Theorem 1: There exists a set of s, homogenous differential equations, $1 \leq s \leq n$ of order r_i, i = 1, 2, ..., s, associated with (1) such that $\sum_{i=1}^{s} r_i = n$.

Proof: Substitute $X = C^{-1}Y$ in (1) so that $\dot{Y} = CAC^{-1}Y$ results where

$$CAC^{-1} = \begin{bmatrix} B_1 & 0 & \ldots & 0 \\ 0 & B_2 & \ldots & 0 \\ & \ldots\ldots & & \\ 0 & 0 & & B_s \end{bmatrix} \quad \text{and } B_i = \begin{bmatrix} 0 & 1 & 0 & \ldots 0 \\ 0 & 0 & 1 & \ldots 0 \\ & \ldots\ldots\ldots & & \\ 0 & 0 & 0 & \ldots 1 \\ a_{r_ir_i} & a_{r_ir_i-2} & & \ldots a_{r_i1} \end{bmatrix}$$

It has been shown$^{(7)}$ that such a transformation exists. For any subset of equations of $\dot{Y} = CAC^{-1}Y$ of the form

$$\dot{Y}_i = B_iY_i \tag{3}$$

where $Y_i' = [y_{i1}, y_{i2}, \ldots, y_{ir_i}]$, calculate r_i-1 successive derivaties of the first row of (3) eliminating each time from the right hand side the first derivatives of $y_{i2}, y_{i3}, \ldots, y_{ir_i}$ by means of the last r_i-1 equations of (3). This process results in

$$[y_{i1} y_{i1}^{(1)} \ldots y_{i1}^{(r_i-1)}]' = [y_{i1} y_{i2} \ldots y_{ir_i}]' \tag{4}$$

Substituting (4) into (3) gives

$$\begin{bmatrix} y_{i1}^{(1)} \\ y_{i1}^{(2)} \\ \vdots \\ y_{i1}^{(r_i-1)} \\ y_{i1}^{(r_i)} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & \ldots\ldots & & \\ 0 & 0 & 0 & \ldots & 1 \\ a_{r_ir_i} & a_{r_ir_i-1} & a_{r_ir_i-2} & \ldots & a_{r_i1} \end{bmatrix} \begin{bmatrix} y_{i1} \\ y_{i1}^{(1)} \\ y_{i1}^{(2)} \\ \vdots \\ y_{i1}^{(r_i-1)} \end{bmatrix} \tag{5}$$

The last row of (5)

$$\frac{d^{r_i} y_{i1}}{dt^{r_i}} = \sum_{j=1}^{r_i} a_{r_i j} \frac{d^{r_i - j} y_{i1}}{dt^{r_i - j}} \tag{6}$$

is associated with (1). For if F_m is a fundamental matrix of (1), by Lemmas* A1 and A2 and the initial part of the proof, there exists non-singular matrices C and D such that a sub-matrix of CF_mD, G_i, is a fundamental matrix of (3). By (4) and (5) each entry in row one of G_i is a solution on I of (6) and the j-row of G_i is the j-1 derivative of the first row for $j = 1, 2, \ldots, r_i$. Since the determinant of G_i is the Wronskian of solutions on I of (6) and since G_i is non-singular, the entries in row one of G_i are a fundamental set of (6).

If $y_1, y_2, \ldots, y_{r_i}$ is a fundamental set of (6), by (5) and (6), r_i linearly independent solutions on I of (3) are $[y_j, y_j^{(1)}, \ldots, y_j^{(r_i-1)}]'$, $j = 1, 2, \ldots, r_i$. This implies that the vectors $Y_j' = [0, \ldots, 0, y_j, y_j^{(1)}, \ldots, y_j^{(r_i-1)}, 0, \ldots, 0]$, $j = 1, 2, \ldots, r_i$ are r_i linearly independent solutions on I of $\dot{Y} = CAC^{-1}Y$. Since $X = C^{-1}Y$, r_i linearly independent solutions on I, of (1), are $C^{-1}Y_j$, $j = 1, 2, \ldots, r_i$. Therefore, (6) is associated with (1). Corresponding to each matrix B_i, of order r_i, $i = 1, 2, \ldots, s$, there exists an associated differential equation of order r_i. Since matrix A is of order n, then $\sum_{i=1}^{s} r_i = n$.

An examination of the characteristic polynomial of (6) shows that if λ_0 is a zero of this polynomial, then λ_0 is also a zero of the characteristic polynomial of (1), $\det[A - \lambda I] = 0$. In addition if $r = n$, by Theorem $2^{(8)}$, $a_{r_i j}$ is $(-1)^{j+1}$ times the sum of the principal minors of order j of A.

The existence Theorem 1 might at first be considered to be of academic interest only. However, in the proof of that theorem, a new method for converting a system of n equations (1) into a higher-order differential equation by transform techniques is given, i.e., $X = C^{-1}Y$. Before the transform technique can be applied, the practical problem of determining a transformation of the type $\mathbf{X} = C^{-1}Y$ which converts the system into an r-order differential equation, must be solved. Consider the system $\dot{\mathbf{X}} = AX$ partitioned as

---

*Lemma A1 and A2 appear in the Appendix.

$$\frac{d}{dt}\begin{bmatrix} X_1 \\ \\ x_n \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ \\ A_{21} & a_{nn} \end{bmatrix}\begin{bmatrix} X_1 \\ \\ x_n \end{bmatrix} \qquad (7)$$

where $X_1' = [x_1, x_2, \ldots, x_{n-1}]$. Take n-1 successive derivatives of $A_{21}X_1 = -a_{nn}x_n + \dot{x}_n$ the last row in (7), eliminate each time the first derivative of the vector X_1 by means of $X_1 = A_{11}X_1 + A_{12}x_n$ the first row in (7). This formulation results in the following matrix form:

$$\begin{bmatrix} A_{21} \\ A_{21}A_{11} \\ \vdots \\ A_{21}A_{11}^{n-2} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} -a_{nn} & 1 & 0 & \ldots 0 & 0 \\ -A_{21}A_{12} & -a_{nn} & 1 & \ldots 0 & 0 \\ \vdots & & & & \\ -A_{21}A_{11}^{n-3}A_{12} & -A_{21}A_{11}^{n-4}A_{12} & -A_{21}A_{11}^{n-5}A_{12} & \ldots -a_{nn} & 1 \end{bmatrix}\begin{bmatrix} x_n \\ x_n^{(1)} \\ \vdots \\ x_n^{(n-1)} \end{bmatrix}$$

This last equation can be written in symbolic form as $BX_1 = PX_d$ where the i-row is

$$A_{21}A_{11}^{i-1}\mathbf{X}_1 = -A_{21}A_{11}^{i-2}A_{12}x_n - \ldots -A_{21}A_{11}^{0}A_{12}x_n^{(i-2)} - a_{nn}x_n^{(i-1)} + x_n^{(i)}$$

$i = 1, 2, \ldots, n-1$. By attaching ones and zeros to this equation form

$$\begin{bmatrix} 0 & 1 \\ \\ B & 0 \end{bmatrix}\begin{bmatrix} X_1 \\ \\ x_n \end{bmatrix} = \begin{bmatrix} U \\ \\ P \end{bmatrix}\begin{bmatrix} X_d \end{bmatrix} = L\begin{bmatrix} X_d \end{bmatrix} \qquad (8)$$

where $U = [1, 0, 0, \ldots, 0]$, $\mathbf{L}' = [U' \; P']$.

If B^{-1} exists, the coefficient matrix on the left side of (8) is non-singular. The coefficient matrix on the right side of (8) is non-singular since it is lower triangular with ones on the main diagonal. Let

$$G = \begin{bmatrix} U \\ \\ P \end{bmatrix}^{-1}\begin{bmatrix} 0 & 1 \\ \\ B & 0 \end{bmatrix} \qquad (9)$$

If the a_{ij} entries of matrix A of the system (1) satisfy the condition $\det(B) \neq 0$ then the matrix G defined by (9) is non-singular. This result is stated by the following lemma:

Lemma 1: If, corresponding to the system (1) the a_{ij} entries of the matrix A satisfy $\det(B) \neq 0$, where the i-row of matrix B is

$$\begin{bmatrix} a_{n1} & a_{n2} & \dots & a_{n,n-1} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} \\ & \dots\dots\dots & & \\ a_{n-1,1} & a_{n-1,2} & \dots & a_{n-1,n-1} \end{bmatrix}^{i-1}$$

$i = 1, 2, \dots, n-1$, then there exists a non-singular matrix G such that

$$X_d = GX \tag{10}$$

where $X'_d = [x_n, x_n^{(1)}, \dots, x_n^{(n-1)}]$.

Theorem 2 establishes that the Transformation (10) converts the state equations into an n-order differential equation which is associated with the system.

Theorem 2: If $\det(B) \neq 0$, where matrix B is defined in the hypothesis of Lemma 1, then there exists an n-order differential equation associated with (1).

Proof: The argument here is similar to that given in the proof of Theorem 1 and therefore has been omitted.

The relationship between the a_{ij} entries of matrix A in (1) and a specified solution to the r-order differential equation (2), associated with (1), is established from the solution properties specified in Definition 1.

Theorem 3: If $y_1, y_2, \dots, y_r$ is a fundamental set for the r-order equation (2) and (2) is associated with the system (1), then there exists an explicit relation between r^2 entries of matrix A and the entries of matric C and vector Y_j, $j = 1, 2, \dots, r$.

Proof: Since (2) is associated with (1) the vectors $C^{-1}Y_j$ where $Y_j = [0, \dots, y_j, y_j^{(1)}, \dots, y_j^{(r-1)}, 0, \dots, 0]'$, $j = 1, 2, \dots, r$ are r linearly independent solutions on I of (1). Let $F_m = [Y_1, Y_2, \dots, Y_r]$ and F_r be the (non-singular) submatrix of F_m containing the columns $[y_j, y_j^{(1)}, \dots, y_j^{(r-1)}]'$, $j = 1, 2, \dots, r$. Since the matrix $C^{-1}F_m$ satisfies (1), $C^{-1}\dot{F}_m = AC^{-1}F_m$. Multiplying the above equation on the right by $[0, F_r^{-1}, 0]$ results in a system of equations in the following form: $C_1^{-1}\dot{F}_r F_r^{-1} = AC_1^{-1}$ where C_1^{-1} is the set of columns of C^{-1} corresponding to the position of the entries in $\dot{F}_r (F_r)$ in $\dot{F}_m (F_m)$. Since the columns of C_1^{-1} are linearly independent, there exists a non-singular submatrix, C_r^{-1}, of C_1^{-1} of order r. Multiplication on the right by C_r results in a system of equations which contains a subset of equations in the form $A_r = C_r^{-1}\dot{F}_r F_r^{-1} C_r$. Since A_r is square and therefore contains r^2 entries of A, the theorem follows.

Corollary 1: With the same hypothesis and $r = n$,

$$A = C^{-1}\dot{F}_m F_m^{-1} C \tag{11}$$

Proof: This is a direct consequence of Theorem 3.

Corollary 2: With the same hypothesis and $r = n$,

$$\dot{F}_m F_m^{-1} = \begin{bmatrix} 0 & 1 & 0 & \dots 0 \\ 0 & 0 & 1 & \dots 0 \\ 0 & 0 & 0 & \dots 1 \\ b_n & b_{n-1} & b_{n-2} & \dots b_1 \end{bmatrix} \tag{12}$$

Proof: Let T represent the companion matrix on the right of (12). Consider the matrix product $T F_m$. If the n equations $b_n y_i + b_{n-1} y_i^{(1)} + \dots + b_1 y_i^{n-1} = y_i^{(n)}$, $i = 1, 2, \dots, n$, are satisfied, then $T F_m = \dot{F}_m$.

The above system of n equations in matrix form is

$$\begin{bmatrix} y_1 & \dot{y}_1 & \dots & y_1^{(n-1)} \\ y_2 & \dot{y}_2 & \dots & y_2^{(n-1)} \\ \cdots & \cdots & \cdots & \cdots \\ y_n & \dot{y}_n & \dots & y_n^{(n-1)} \end{bmatrix} \begin{bmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \end{bmatrix} = \begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix}$$

Since F_m is non-singular and the matrix on the left of the above equation is F_m', the lemma follows:

Corollary 1 establishes a relationship (1) for determining the a_{ij} entries of the matrix A in (1) in terms of one specified component

$x_i(t)$ of the vector $X(t)$ of (1). When Theorem 2 is applied to Corollary 1 it is determined that if the a_{ij} entries in matrix A satisfy det (B) $\neq$ 0 then the vector component $x_n(t)$ of $X(t)$ in the system (1) can be specified as consisting of n linearly independent parts. This specification in turn restricts the a_{ij} entries of matrix A by

$$A = G^{-1} \dot{F}_m F_m^{-1} G \tag{13}$$

This relation is useful for purposes of calculation while (11) is not, although matrix C has been shown to exist,(7) no well-defined method of calculating C is given.

An important consequence of Corollary 2 is that the matrix products CAC^{-1} and GAG^{-1} have the form of companion matrices. This implies that the characteristic polynomial of A, det $[A - \lambda I]$ must be equal to the minimum polynomial of A if det (B) $\neq$ 0.

Gantmacher(9) shows that the matrix C in the similarity transformation CAC^{-1}, which produces the companion matrix, is not unique. Faddeeva(10) presents a method developed by Danilevsky which brings a matrix A into companion matrix form by means of (n-1) similarity transformations. The primary difference in the method for obtaining the companion matrix in this paper and other methods is that here it is obtained under specified conditions det (B) $\neq$ 0 on the a_{ij} entries of matrix A by a particular matrix product GAG^{-1}.

The fact that GAG^{-1} has the form of a companion matrix is applied in the following theorem.

Theorem 4: If in the system (1), det (B) $\neq$ 0 where matrix B is defined in Lemma 1, then the r-order equation (2) where a_j is $(-1)^{j+1}$ times the sum of the principal minors of order j of A, is associated with (1).

Proof: This follows from Theorem 2 and Corollary 2.

Theorem 4 provides a simple, yet effective procedure for establishing a mathematical relationship between the a_{ij} entries of the matrix A in the system (1) and the a_j entries in the n-order equation

(2). If the matrix A is given and det (B) $\neq$ 0, or if matrix A contains arbitrary entries and det (B) $\neq$ 0 is specified, then the mathematical relationships are given in the theorem.

III. NONHOMOGENOUS STATE EQUATIONS AND ASSOCIATED HIGHER ORDER DIFFERENTIAL EQUATIONS

Properties parallel to those in Section II are developed in this section for nonhomogenous state equations. The precise relationship between solutions of nonhomogenous state equations (14) and higher order differential equations obtained from (14) is stated in the following definition and proven in the lemma and theorem which follow the definition.

Definition 2: Consider the system of n-nonhomogenous equations,

$$\dot{X} = AX + Q(t) \tag{14}$$

where $A = [a_{ij}]$, and a_{ij} is a constant, $\mathbf{X}' = [x_1, x_2, \dots, x_n]$, $Q'(t) = [q_1(t), \dots, q_n(t)]$, $q_i^{(n)}(t)$, $i = 1, 2, \dots, n$ is continuous on the open interval I defined by $I = [t : t_1 < t < t_2]$, where t_1 and t_2 are constants.

An r-order nonhomogenous differential equation

$$\frac{d^r y}{dt^r} = \sum_{i=1}^{r} a_i \frac{d^{r-i} y}{dt^{r-i}} + F(t) \tag{15}$$

is associated with (14) if:

(a) The homogenous part of (15) is associated (with non-singular matrices C and D and row i) with the homogenous part of (14) and

(b) 1. For $\mathbf{X}'(t) = [x_1(t), x_2(t), \dots, x_n(t)]$ the solution of (14) on I such that $\mathbf{X}(t_o) = 0$, t_o on I, then row i of $C\mathbf{X}(t)$ is the solution of (15) on I which is zero at t_o.

2. For y(t) the solution of (15) on I such that $y^{(j)}(t_o) = 0$, for $j = 0, 1, \dots, r-1$, t_o on I, and

$$F(t) = -\sum_{i=1}^{r-1} a_i \sum_{j=0}^{r-i-1} P_j^{r-i+1}(D)\, f_{r-i-j}(t) + \sum_{j=0}^{r-1} P_j^{r}(D)\, f_{r-j}(t) \tag{16}$$

where $P_j^k(D) = D^j + \ell_{k,k-1} D^{j-1} + \ell_{k,k-2} D^{j-2} + \dots + \ell_{k,k-j}$,

$D^j = \frac{d^j}{dt^j}$, $P_o^k(D) = 1$ and $\sum_{j=0}^{k-2} P_j^k(D)\, f_{k-1-j}(t_o) = 0$, $k = 2, 3, \ldots, r$, then $C^{-1}[Y_s(t) - L^{-1}H(t)]$ is the solution of (14) which, when evaluated at t_o, is the zero vector, where $L^{-1} = [\ell_{ij}]$ and $Y_s'(t) = [0, 0, \ldots, 0, y(t), \ldots, y^{(r-1)}(t), 0, \ldots, 0]$, $H'(t) = [0, 0, \ldots, 0, f_1(t), f_1^{(1)}(t) + f_2(t), \ldots, f_1^{(r-2)}(t) + f_2^{(r-3)}(t) + \ldots + f_{r-1}(t), 0, \ldots, 0]$.

Definition 3: An r-order homogenous differential equation (1) is associated with (14) if (a) and (b)1. of Definition 2 are satisfied.

Lemma 2: Suppose matrix A of the system (14) is

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & \cdots\cdots & & \\ 0 & 0 & 0 & \ldots & 1 \\ a_n & a_{n-1} & a_{n-2} & \ldots & a_1 \end{bmatrix}$$

Then there exists an n-order differential equation, associated with the system (14).

Proof: This follows by calculating the n-1 successive derivatives of the first row of (14) eliminating each time from the right hand side, the first derivatives of $x_2, x_3, \ldots, x_n$ by means of the last n-1 original equations. This determines(11) an n-order differential equation associated with (14).

Theorem 5: There exists a set of s differential equations, $1 \leq s \leq n$, of order r_i, $i = 1, 2, \ldots, s$, associated with the system (14) such that $\sum_{i=1}^{s} r_i = n$.

Proof: This follows using the transformation and argument used in the proof of Theorem 1, Lemma 2, and by an argument similar to that used in the proof of Lemma 2.

The following lemma gives the detailed form of the transformation of the type $X = C^{-1}Y$ such that the higher order differential equation associated with (14) is of order n.

Lemma 3: If corresponding to the system (14), det (B) $\neq 0$, then there exists a vector $Q_d(t)$ and non-singular matrices G and L such that

$$X_d = GX + L^{-1}Q_d(t) \tag{17}$$

where matrix B is defined in the hypothesis of Lemma 1, matrices G and L are given in (9) and (8) respectively, $Q'_d(t) = [0 \; Q'_2(t)]$, $Q'_2(t) = [0, -q_n(t), -q_n^{(1)}(t) + A_{21}A_{11}^{o} Q_1(t), -q_n^{(2)}(t) + A_{21}A_{11}^{o} Q_1^{(1)}(t) + A_{21}A_{11}^{1}Q_1(t), \ldots, -q_n^{(n-2)}(t) + A_{21}A_{11}^{o} Q^{(n-3)} + \ldots + A_{21}A_{11}^{(n-3)}Q_1(t)]$, $Q_1(t)$ is the vector consisting of the first (n-1) entries in vector Q and $X'_d = [x_n x_n^{(1)} \ldots x_n^{(n-1)}]$.

Proof: This follows by a calculation similar to that given in Section 1, (7)-(9).

Theorem 6: If det (B) ≠ 0, where matrix B is defined in the hypothesis of Lemma 1, then there exists an n-order differential equation associated with the system (14).

Proof: The arguments used to prove this theorem are similar to those appearing in the proof of Theorem 1 and Lemma 2 and therefore have been omitted.

The above results not only offer a new method of formulating an n-order differential equation (last row of system of equations obtained from substituting (17) into (14)), but more important, a closed form relationship $X = G^{-1}[X_d - L^{-1}Q_d(t)]$ linking the solution of an n-order differential equation to the solution of the state equations.

The details of the proof of Theorem 6 indicate that the following n-order differential equation is associated with the state equations (14).

$$\frac{d^n}{dt^n} x_n = \sum_{i=1}^{n} a_i \frac{d^{n-i}}{dt^{n-i}} x_n - \sum_{i=1}^{n-1} a_i \sum_{j=0}^{n-i+1} P_j^{n-i+1}(D) f_{1,n-i-j}(t) + \sum_{j=0}^{n-1} P_j^n(D) f_{1,n-j}(t) \tag{18}$$

where $P_j^k(D)$ is given after (16), $L^{-1} = [\ell_{ij}]$ and $f_{11}(t) = q_n(t)$, $f_{1i}(t) = A_{21}A_{11}^{i-2}Q_1(t)$, $i = 2, 3, \ldots, n$. The a_i, $i = 1, 2, \ldots, n$ components in (18) are (Thm. 4) $(-1)^{i+1}$ times the sum of the principal minors of order i of matrix A in the state equations (14). In addition, the nonhomogenous part of the state equations is found to be $Q'(t) = [f_1(t), f_2(t), \ldots, f_n(t)]\ (G^{-1}L^{-1})'$ where $f_1(t), f_2(t), \ldots f_n(t)$ are the components in the nonhomogenous part $F(t)$ of the n-order equation (15). In the following theorem new relations between the coefficient matrix A and the vector $Q(t)$ in the state equations (14) and the solution $y(t)$ and the nonhomogenous part $F(t)$ of an n-order differential equation (15) associated with (14) are given.

Theorem 7: Suppose $y_1(t), y_2(t), \ldots, y_n(t)$ is a fundamental set of the homogenous part of the n-order equation (15), $y(t)$ is the solution on I: $[t\,|>t_o$ of (15) for $r = n$ such that $y^{(j)}(t_o) = 0$, t_o on I, $j = 0, 1, \ldots, n-1$ and (15) ($r = n$) is associated with (14), then

$$A = C^{-1}\dot{F}_m(t)\,F_m^{-1}(t)C, \qquad Q(t) = F_m(t)\,\frac{d}{dt}\,(F_m^{-1}(t)\,Y(t))$$

where $F_m(t) = [f_{ij}(t)]$, $f_{ij}(t) = y_j^{(i-1)}(t)$, $i, j = 1, 2, \ldots, n$ and $Y(t) = C^{-1}[Y_s(t) - L^{-1}H(t)]$.

Proof: By hypothesis, and Corollary 1, the first part of the conclusion follows.

If $F_m(t)$ is a fundamental matrix for $\dot{X} = AX$ then, by Theorem 3.1$^{(6)}$ $Y(t) = F_m(t)\int_{t_o}^{t} F_m^{-1}(s)Q(s)ds$ t on I, is that solution of (14) satisfying $Y(t_o) = 0$. Application of "Leibnitz Rule$^{(12)}$ to $\frac{d}{dt}(F_m^{-1}(t)Y(t)) = \frac{d}{dt}(\int_{t_o}^{t} F_m^{-1}(s)Q(s)ds)$ results in n relations, $Q(t) = F_m(t)\,\frac{d}{dt}\,(F_m^{-1}(t)Y(t))$. Since the n-order equation (15) is associated with the system (14) then $Y(t) = C^{-1}[Y_s(t) - L^{-1}H(t)]$ is the solution of (14) which is zero when evaluated at t_o, t_o on I. This implies the second part of the conclusion.

In Theorem 7, if $y(t)$ is replaced by $x_n(t)$ and the restriction det (B) $\neq 0$ (Lemma 1) is added to the hypothesis then the relationships,

$$A = G^{-1}\dot{F}_m F_m^{-1}G, \qquad Q(t) = F_m(t)\,\frac{d}{dt}\,(F_m^{-1}(t)X(t))$$

are obtained. The solution of (14) is given in this restricted situation i.e., det (B) $\neq 0$, by the vector $X(t) = G^{-1}[X_d(t) - L^{-1}H(t)]$ where $X_d'(t) = [x_n(t), x_n^{(1)}(t), \ldots, x_n^{(n-1)}(t)]$, matrices G and L are given in

(9) and (8) respectively and H(t) is determined as specified in Definition 2 where $r = n$.

IV. DESIGN PROCEDURES

One approach to the design problem in the time domain is to view the system performance as one specified component $x_i(t) = y(t)$ of the solution vector X(t) of the state equations. Then by the mathematical techniques developed in the foregoing, determine the relationships including initial conditions (design equations) which must be satisfied between the parameters in the state equations and the specified component.

For the case of homogenous state equations (1), the specified component must have the following form (general solution of (2)): $y(t) = \sum_{i=1}^{k} P_{m_i-1}(t)\, e^{\lambda_i t}$ where $P_{m_i-1}(t)$ is a polynomial in t of degree $m_i - 1$, λ_i, $i = 1, 2, \ldots, k$ are the distinct zeros of $\lambda^n - \sum_{i=1}^{n} a_i \lambda^{n-i} = 0$ each of multiplicity m_i and $\sum_{i=1}^{k} m_i = n$.

For nonhomogenous state equations (14), the specified component must have the form $y(t) = y_h(t) + y_p(t)$ where $y_h(t)$ satisfies (2) and $y_p(t) = \sum_{i=1}^{n} \emptyset_i(t) \int_{t_0}^{t} \frac{w_k(s)}{w(s)} F(s)\, ds$ where $\emptyset_i(t)$, $i = 1, 2, \ldots, n$ is a fundamental set for (2), W (s) is the Wronskian of (2) and $W_k(s)$ is the determinant obtained from the Wronskian W (s) by replacing the k-column by $[0\ 0 \ldots 0\ 1]'$.

Methods 1 and 2 which follow present general procedures for determining the interrelations for homogenous state equations.

Method 1: (a) Apply Theorem A1 of Appendix to $y(t) = \sum_{i=1}^{k} P_{m_i-1}(t)e^{\lambda_i t}$ and thereby determine the coefficients a_j, $j = 1, 2, \ldots, n$ of the n-order equation (2).

(b) Relate a_j, $j = 1, 2, \ldots, n$ to the entries a_{ij} of matrix A in the state equations (1) by application of Theorem 4.

(c) Relate the initial condition $y^{(j)}(t_0) = C_{j+1}$, $j = 0, 1, \ldots, n-1$, to the initial condition on the state equations (1) by (10).

Method 2: (a) From the matrix

$$F_m = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ y_1^{(1)} & y_2^{(1)} & \cdots & y_n^{(1)} \\ & \cdots\cdots & & \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{bmatrix}$$

from a fundamental set $y_1, y_2, \ldots, y_n$, of the n-order equation (2).

(b) Determine the matrix A of the system (1) as

$$A = G^{-1} \dot{F}_m F_m^{-1} G.$$

(c) Determine the initial conditions for the system (1) as in (c) of Method 1.

Using the results of the foregoing theorems, two methods of determining the interrelations for nonhomogenous equations are determined.(11) The resulting equations are the design equations (including initial conditions) for the system.

APPENDIX

The following lemmas are used in the proof of Theorem 1.

Lemma A1: If F_m is a fundamental matrix of the system (1) and $Y = CX$, where C^{-1} exists, then CF_m is a fundamental matrix of $\dot{Y} = CAC^{-1}Y$.

Proof: Let f_j, $j = 1, 2, \ldots, n$ be the j-column of F_m. Since f_j is a solution on I of the system (1) and is obtained by substituting $X = C^{-1}Y$ into (1), Cf_j is a solution of $\dot{Y} = CAC^{-1}Y$ on I. Since C and F_m are non-singular by hypothesis, CF_m is non-singular. The lemma follows.

Lemma A2: Suppose CAC^{-1} of Lemma A1 is

$$CAC^{-1} = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ 0 & 0 & \cdots & B_s \end{bmatrix}$$

where B_i is of order r_i, $i = 1, 2, \ldots, s$. There exists a non-singular matrix D such that a fundamental matrix of $\dot{Y} = B_i Y_i$, $i = 1, 2, \ldots, s$ where $Y' = [Y_1', Y_2', \ldots, Y_s']$, is the submatrix of $CF_m D$ consisting of the entries in rows $\sum_{j=1}^{i-1} r_j+1, \sum_{j=1}^{i-1} r_j+2, \ldots, \sum_{j=1}^{i} r_j$ and columns $\sum_{j=1}^{i-1} r_j+1, \sum_{j=1}^{i-1} r_j+2, \ldots, \sum_{j=1}^{i} r_j$.

Proof: This lemma follows from Lemma A1 and Theorem 2.3.(5)

The following theorems are used to relate the coefficients of the higher-order differential equation to the parameters of the corresponding solution.

Theorem A1: 1. If a solution on I of the n-order equation (2) is $y(t) = \sum_{i=1}^{k} P_{m_i-1}(t)\, e^{\lambda_i t}$, then $a_1 = +(r_1 + r_2 + \ldots + r_n)$

$$a_2 = -(r_1 r_2 + \ldots + r_1 r_n + r_2 r_3 + \ldots + r_2 r_n + \ldots + r_{n-1} r_n)$$

$$\vdots$$

$$a_n = -(-1)^n (r_1 r_2 \ldots r_n) \text{ and}$$

$$r_i = \lambda_j, \; i = \sum_{p=1}^{j-1} m_p+1, \sum_{p=1}^{j-1} m_p+2, \ldots, \sum_{p=1}^{j} m_p, \; j = 1, 2, \ldots, k.$$

2. If a solution on I of the n-order equation (2) is $y(t) = \sum_{i=1}^{r} c_i e^{\lambda_i t}$, $c_i \neq 0$, $\lambda_i \neq 0$ and λ_i's are distince for all i, then r coefficients of (2) are given by the vector $a = V^{-1} L$ where $a' = [b_j b_{j+1} \ldots b_{j+r-1}]$, $L' = [-\lambda_1^n - \sum_s b_i \lambda_1^{n-i} \ldots -\lambda_r^n - \sum_s b_i \lambda_r^{n-i}]$, $\sum_s$ is the sum over all i, $i \neq j, j+1, \ldots, j+r-1$, $a_j = -b_j$ and

$$V = \begin{bmatrix} \lambda_1^{n-j} & \lambda_1^{n-j-1} & \ldots & \lambda_1^{n-j-r+1} \\ \lambda_2^{n-j} & \lambda_2^{n-j-1} & \ldots & \lambda_2^{n-j-r+1} \\ & \ldots\ldots\ldots & & \\ \lambda_r^{n-j} & \lambda_r^{n-j-1} & & \lambda_r^{n-j-r+1} \end{bmatrix}$$

Proof: Since λ_i are zeros of $L(\lambda) = \lambda^n + \sum_{j=1}^{n} b_j \lambda^{n-j}$ of order m_i, substituting and grouping coefficients of terms of the same order proves the conclusion of 1. The conclusion of 2 follows as the solution of the system of equations $L(\lambda_i) = \lambda_i^n + \sum_{j=1}^{n} b_n \lambda_i^{n-j} = 0$, $i = 1, 2, \ldots, r$.

Footnotes

1. L. A. Zadeh, "Time-Varying Networks I," Proc. IRE, Vol. 49, (October 1961) pp. 1488-1503.

2. F. Murray and K. Miller, Existence Theorems, New York: University Press, 1954, P. 126.

3. Moulton (4) presents, for example, a system of three differential equations which converts into a second-order equation in any component, x_i.

4. F. R. Moulton, Differential Equations, Dover, 1958, P. 9.

5. E. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill Company, Inc., 1955, P. 21.

6. W. Hurewicz, Lectures on Ordinary Differential Equations, John Wiley and Sons, Inc., 1958, P. 33.

7. H. W. Turnbull and A. C. Aitken, An Introduction to the Theory of Canonical Matrices, Dover, 1961, P. 49.

8. F. E. Hohn, Elementary Matrix Algebra, MacMillan Company, 1958, P. 215.

9. F. R. Gantmacher, Theory of Matrices, Vol. 1, New York: Chelsea Publishing Company, 1959, P. 159.

10. V. N. Faddeeva, Computational Methods of Linear Algebra, Dover, 1959, P. 201.

11. R. H. Rogers, On Properties of a Class of Systems of Differential Equations and Corresponding Higher Order Differential Equations, Ph.D. dissertation, Michigan State University, East Lansing, Michigan, 1963.

ON THE DIFFERING ABILITIES OF RL STRUCTURES TO REALIZE NATURAL FREQUENCIES*

H. B. Lee⁺

I. INTRODUCTION

It is well known that the natural frequencies of RL structures must be real and negative if the element values of the structure are positive. Moreover, it is known that by using specific RL structures such as the Foster and the Cauer structures, one can realize any prescribed set of such natural frequencies while employing only positive elements.

Beyond the foregoing results, little seems to be known about the natural frequency realizing abilities of RL structures. In particular, little appears to have been done on the question of identifying other RL structures which have the same ability for realizing natural frequencies as the Foster and Cauer structures.

In the present paper we attempt to cast some light on these matters. Specifically we study a restricted class of RL structures with the object of distinguishing between

1. the member structures that approximately equal the Foster and Cauer structures in natural frequency realizing ability

and

2. the member structures that are markedly inferior to the Foster and Cauer structures in natural frequency realizing ability.

Our main result consists of a necessary and sufficient condition for the member structures to belong to the former category. The condition in question is remarkably simple and suggests a corresponding condition as obtaining for the entire class of RL structures.

The results of the paper, besides being of interest in themselves, are useful in estimating the synthesis potential of unconventional networks.

II. NOMENCLATURE

The specific RL structures which we consider are those in which the resistors form a single tree, and the inductors form a single tree. We refer to these structures as *complementary tree structures.*

To indicate the sizes of different complementary tree structures, we employ the notion of *complexity* due to Reza.[1] Because complementary tree structures contain neither cut sets of resistance nor cut sets of inductance, their complexities simply equal the numbers of their independent nodes.

Complementary tree structures of complexities two, three, four and five are shown in Figures 1, 2, 3 and 4 respectively.[2]

To simplify matters, we say that a complementary tree structure is *good* if its natural frequency realizing ability approximates that of the Foster and the Cauer structures; if this is not the case, then we say that it is *bad*. More precisely, we adhere to the following definition.

Definition 1: If the elements of an RL structure of complexity n can be adjusted to positive values such that the structure realizes, within an arbitrarily small uniform tolerance ϵ, any prescribed set of natural frequencies

$$s = -\sigma_1, -\sigma_2, \ldots, -\sigma_n \; (\sigma_n > \sigma_{n-1} \cdots > \sigma_1 > 0),$$

then the structure is *good*; otherwise the structure is *bad*.

The problem which we wish to consider is easily stated with the aid of the foregoing terminology. The problem is that of separating the complementary tree structures into good structures and bad structures.

III. THE SOLUTION

To solve the problem we study the properties of a general complementary tree structure of complexity i. The solution reveals itself as the study progresses.

We denote the general complementary tree structure by S_i. Throughout we assume that the elements of S_i can take on only positive values. Single resistors and single inductors of S_i are denoted by small r's and ℓ's. Each resistor of S_i possesses a single inductive path between its terminals provided by the inductive tree of S_i; the inductance in such a path is denoted by a capital L. Similarly, each inductor of S_i possesses a unique resistive path between its terminals provided by the resistive tree of S_i; the resistance in such a path is denoted by a capital R.

We begin by proving two preliminary theorems about S_i.

Theorem 1: If r is any one of S_i's resistors, and L is the inductive path associated with r, then the following relationship obtains for every set of element values

$$|s_{max}| > \frac{r}{L} > |s_{min}| \quad , \tag{1}$$

where s_{max} and s_{min} are, magnitudewise, the largest and smallest natural frequencies of S_i.

Proof: Let the elements of S_i be adjusted to arbitrary values. Let Z denote the impedance into which r looks. The σ-axis plot of Z has the character shown in Figure 5 (note that the intermediate poles of Z have been omitted). The zero at $s = 0$ is due to the inductive path L, and for this reason $\left.\frac{dZ}{ds}\right|_{s=0} = L$. The inductive behavior of Z at $s = \infty$ is due to the fact that r cannot look into a resistive path (if r did look into a resistive path, then the path and r would form a resistive loop, contradicting the fact that the resistors of S_i form a tree). The natural frequencies of S_i are given by the intersections between the $Z(\sigma)$ vs. σ curve and the line $Z = -r$, as shown in Figure 6. From the familiar properties of the $Z(\sigma)$ vs. σ curve, it is clear that

$$|s_{max}| > \frac{r}{L} > |s_{min}| \quad .$$

Q.E.D.

The second preliminary theorem is the complement of Theorem 1. It reads

Theorem 2: If ℓ is any one of S_i's inductors, and R is the resistive path associated with ℓ, then the following relationship obtains for every set of element values

$$|s_{max}| > \frac{R}{\ell} > |s_{min}| \quad , \tag{2}$$

where s_{max} and s_{min} are, magnitudewise, the largest and smallest natural frequencies of S_i.

The proof of this theorem is omitted, since it differs from the proof of Theorem 1 only by an interchange of resistive and inductive behavior.

We next demonstrate a very interesting property of S_i.

Theorem 3: If each loop of S_i contains three or more branches, then the following relationship obtains for every set of element values

$$\left|\frac{s_{max}}{s_{min}}\right| > 4 \quad . \tag{3}$$

Proof: Assume that every loop of S_i contains three or more branches, and let the elements of S_i be adjusted to any set of values whatever.

Consider the loop formed by the smallest of S_i's resistors and its associated inductive path. Let r_1 denote the resistor in question, and L_1 its associated inductive path. From Theorem 1 it follows that

$$r_1 > L_1 \cdot |s_{min}| \quad . \tag{4}$$

By hypothesis, the loop consisting of r_1 and L_1 contains three or more branches. Thus the inductive path L_1 must contain two or more branches. That is, if $n_{\ell 1}$ denotes the number of inductances contained in L_1, then

$$n_{\ell 1} \geqq 2 \quad . \tag{5}$$

Moreover, if ℓ_1 denotes the smallest inductance in L_1, the following relationship obtains

$$\frac{L_1}{n_{\ell 1}} \geqq \ell_1$$

or

$$L_1 \geqq n_{\ell 1} \ell_1 \quad , \tag{6}$$

since the smallest inductance in L_1 cannot exceed the average inductance in L_1.

Conditions (4) and (6) together show that

$$r_1 > n_{\ell 1} \ell_1 |s_{min}| \quad . \tag{7}$$

Consider next the loop formed by ℓ_1 and its associated resistive path. Let R_1 denote the resistive path in question. From Theorem 2 it is evident that

$$\ell_1 > \frac{R_1}{|s_{max}|} \quad . \tag{8}$$

Application to the resistive path R_1 of the same reasoning which led to conditions (5) and (6) for the inductive path L_1, indicates that

$$n_{r1} \geqq 2 \tag{9}$$

and

$$R_1 \geqq n_{r1} r_2 \ , \tag{10}$$

where n_{r1} denotes the number of resistors in R_1, and r_2 denotes the smallest resistor in R_1.

Conditions (8) and (10) together show that

$$\ell_1 > \frac{n_{r1}}{|s_{max}|} \cdot r_2 \quad . \tag{11}$$

Substitution of (11) into (7) yields

$$r_1 > n_{\ell 1} n_{r1} \cdot \frac{|s_{min}|}{|s_{max}|} \cdot r_2$$

or

$$\frac{|s_{max}|}{|s_{min}|} > n_{\ell 1} n_{r1} \left(\frac{r_2}{r_1}\right) . \tag{12}$$

Because r_1 represents the smallest of S's resistors,

$$\frac{r_2}{r_1} \geqq 1 \quad .$$

This fact, together with (5) and (9) shows

$$\frac{|s_{max}|}{|s_{min}|} > 4 \quad . \tag{13}$$

Q.E.D.

The foregoing theorem casts a great deal of light upon our problem, for it establishes the bad character of many complementary tree structures. In particular, the theorem shows that every complementary tree structure which contains three or more branches in every loop cannot realize natural frequency sets having $\left|\frac{s_{max}}{s_{min}}\right| < 4$ (every set of element values leads to the result $\left|\frac{s_{max}}{s_{min}}\right| > 4$). Thus the theorem shows that all such structures are bad.[3]

The following theorem is an important consequence of Theorem 3.

Theorem 4: If S_i can realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$, then S_i contains the loop shown in Figure 7.

Proof: Assume that S_i can realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$, and let S_i's elements be adjusted to any set of values which produces this result.

Theorem 3 cannot apply to S_i since $\left|\frac{s_{max}}{s_{min}}\right| < 4$. Therefore at least one of S_i's loops must involve less than three branches. That is, S_i must contain a two-branch loop. Since loops of two resistors or two inductors are prohibited (the resistors form a tree, and so do the inductors), the contained loop must be that shown in Figure 7.

Q.E.D.

The following corollary is an obvious consequence of Theorem 4.

Corollary: If S_i is good, then S_i contains the loop shown in Figure 7.

The good complementary tree structures are pinned down further now. To accomplish this, we assume that

i) S_i contains one or more loops of the type shown in Figure 7;

ii) S_{i-1} is the structure obtained from S_i by shorting out any one loop of the type shown in Figure 7

and prove the following theorems about S_i and S_{i-1}.

Theorem 5: S_{i-1} is a complementary tree structure of complexity i-1.

Proof: The following considerations show that the resistance subgraph of S_{i-1} forms a tree.

A. The resistance subgraph of S_i forms a tree.

B. The resistance subgraph of S_{i-1} is obtained from that of S_i by short circuiting r (i.e. the resistive branch of Figure 7).

C. Short circuiting any branch of the tree of A produces a simpler resistance tree.

Analagous considerations show that the inductance subgraph of S_{i-1} forms a tree. Thus S_{i-1} is a complementary tree structure.

That the complexity of S_{i-1} equals $i-1$ follows from the fact that S_{i-1} has one less independent node than S_i.

Q.E.D.

Theorem 6: For every set of element values, the ratio $\left|\frac{s_{max}}{s_{min}}\right|$ for S_{i-1} is smaller than the corresponding ratio for S_i.

Proof: Let the elements of S_i be adjusted to arbitrary values. Let Z denote the impedance into which the loop of Figure 7 looks. The behavior of Z along the σ-axis of the s-plane will be as shown in Figure 8. The non-zero behavior at $s = 0$ is due to the fact that the loop of Figure 7 does not look into an inductive path (the loop of Figure 7 cannot look into an inductive path, for if it did, then the path and ℓ would form an inductive loop, which is not allowed). The inductive behavior at $s = \infty$ is due to the fact that the loop of Figure 7 does not look into a resistive path. The extreme finite zeros of $Z(s)$ correspond to the s_{max} and s_{min} for S_{i-1}. The impedance z of the loop of Figure 7 will have the σ-axis behavior shown in Figure 9. The natural frequencies of S_i will occur when $Z = -z$, as shown in Figure 10. Clearly

$$\left|s_{max}\right|_{S_{i-1}} < \left|s_{max}\right|_{S_i}$$

and

$$\left|s_{min}\right|_{S_{i-1}} > \left|s_{min}\right|_{S_i}$$

so that

$$\frac{\left|s_{max}\right|_{S_{i-1}}}{\left|s_{min}\right|_{S_{i-1}}} < \frac{\left|s_{max}\right|_{S_i}}{\left|s_{min}\right|_{S_i}}$$

Q.E.D.

Theorem 7: If S_i can realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$, then S_{i-1} also can realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$.

Proof: Obvious from Theorem 6.

With the help of the foregoing theorems we now can proceed to the main result of the paper. Thus assume that S_n is a good complementary tree structure of complexity n. Clearly S_n must be able to realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$. This fact and Theorem 4 show that S_n contains the loop shown in Figure 7. Theorem 5 shows that when this loop is short circuited, a simpler complementary tree structure S_{n-1} results. Theorem 7 shows that S_{n-1} also must be able to realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$. Application of the same line of argument to S_{n-1} shows that: S_{n-1} contains a loop of the type shown in Figure 7; when this loop is short circuited a simpler complementary tree structure S_{n-2} results; S_{n-2} also must be able to realize natural frequency sets for which $\left|\frac{s_{max}}{s_{min}}\right| < 4$. By continuing in this manner it becomes clear that it is possible to reduce S_n to a complementary tree structure of complexity zero (i.e. a single node) by successively short circuiting loops of the type shown in Figure 7.

At this point it is convenient to make the following definition.

Definition 2: Henceforth, any complementary tree structure which can be reduced to a single node, by successively short circuiting loops of the type shown in Figure 7, will be called a shunt reducible complementary tree structure.

With this definition, our result may be stated as follows.

Theorem 8: A necessary condition for S_n to be good is that S_n be shunt reducible.

Not only is the property of shunt reducibility necessary for a complementary tree structure to be good, but it also is sufficient, for we can prove the following theorem.

Theorem 9: If S_n is a shunt reducible complementary tree structure of complexity n, then the elements of S_n can be adjusted to positive values so as to make S_n realize any prescribed set of natural frequencies, $s = -\sigma_1, -\sigma_2, \ldots, -\sigma_n$ $(\sigma_n > \sigma_{n-1} \cdots > \sigma_1 > 0)$ with arbitrarily small error.

Proof: Assume that S_n is a shunt reducible complementary tree structure of complexity n. In addition, assume that S_n is required to realize natural frequencies $s = -\sigma_1, -\sigma_2, \cdots, -\sigma_n$ $(\sigma_n > \sigma_{n-1} \cdots > \sigma_1 > 0)$.

Because S_n can be reduced to a single node by shorting out structures of the type shown in Figure 7, S_n can be constructed by reversing this procedure. That is, S_n can be constructed by splitting a single node, inserting a structure of the type shown in Figure 7 between the two halves of the node, then breaking into a short circuit of the resulting structure, inserting another structure of the type shown in Figure 7, etc.[4]

To realize the frequencies $s = -\sigma_1, -\sigma_2, \cdots, -\sigma_n$ with arbitrarily small error, it is only necessary to imagine S_n as having been "built up" in the above manner, and adjust the elements of S_n to positive values such that

i) the initial parallel resistor-inductor pair realizes the natural frequency $s = -\sigma_1$ (by itself);

ii) the second resistor-inductor pair realizes the natural frequency $s = -\sigma_2$ (before insertion), and has an impedance level far below that of the first resistor-inductor pair;

iii) the third resistor-inductor pair realizes the natural frequency $s = -\sigma_3$ (before insertion), and has an impedance level far below those of the preceding resistor-inductor pairs;

etc.

The successively inserted resistor-inductor pairs will interact only slightly because of the differences in their impedance levels. Thus the resultant structure will realize natural frequencies which are very close to the required values $s = -\sigma_1, -\sigma_2, \cdots, -\sigma_n$. If the errors are too large, they can be made as small as necessary by repeating the procedure and further separating the impedance levels of successive resistor-inductor pairs.[5]

Q.E.D.

Thus we have obtained a complete solution to the problem of separating the good complementary tree structures from the bad complementary tree structures. The good complementary tree structures are those which are shunt reducible; the bad complementary tree structures are those which are not.

Examples of good complementary tree structures are given in Figures 1 and 3; some bad complementary tree structures are shown in Figures 2, 4 and 12.

IV. THE DUAL SOLUTION

By means of an argument dual to that of the preceding section, one can obtain an alternate solution to the problem of distinguishing the good complementary tree structures from the bad. Because the reader doubtless can carry through the dual argument himself, we omit the details and simply state the results.

One finds that a good complementary tree structure must contain a two branch cut set of the type shown in Figure 13. In addition, when the cut set of Figure 13 is erased, there must exist another such cut set among the resulting structures. When this cut set is erased, there must exist another such cut set among the structures then resulting, etc.

The foregoing process is summarized in the following definition.

Definition 3: If a complementary tree structure of complexity n can be reduced to a set of $n+1$ isolated nodes by successively erasing cut sets of the type shown in Figure 13, then the structure is said to be series reducible.

Thus the dual solution asserts that a complementary tree structure is good if and only if it is series reducible.

It is readily verified that the good structures of Figures 1 and 3 are series reducible, while the bad structures of Figures 2, 4 and 12 are not.

A little reflection shows that the properties of series reducibility and shunt reducibility are equivalent. Indeed, if one open circuits the branches of a complementary tree structure in an order opposite to that in which one short circuits them when performing a shunt reduction, then the result is a series reduction. Similarly, if one short circuits branches in an order opposite to that in which one open circuits them when performing a series reduction, then the result is a shunt reduction.

V. CONJECTURES

The fact that the complementary tree structures can be separated so easily into good structures and bad ones raises the question: What is the rule for separating the entire class of RL structures into good structures and bad ones?

On the basis of the preceding study we should like to conjecture an answer to this question. Before doing so, however, it is necessary to introduce the following definitions.

Definition 4: A mixed loop refers to any loop of an RL structure that contains both resistors and inductors.

Definition 5: If the act of short circuiting a mixed loop of an RL structure reduces the complexity of the structure by unity, then the loop is called a simple loop.

Definition 6: If an RL structure of complexity n can be reduced to a structure of complexity zero by successively short circuiting simple loops, then the structure is said to be a loopwise reducible structure.

Definition 7: A mixed cut set refers to any cut set of an RL structure that contains both resistors and inductors.

Definition 8: If the act of open circuiting a mixed cut set of an RL structure reduces the complexity of the structure by unity, then the cut set is called a simple cut set.

Definition 9: If an RL structure of complexity n can be reduced to a structure of complexity zero by successively open circuiting simple cut sets, then the structure is said to be a cut-setwise reducible structure.

Our conjectures are as follows:

Conjecture I. A general RL structure is good if and only if it is loopwise reducible.

Conjecture II. A general RL structure is good if and only if it is cut-setwise reducible.

The following facts recommend these conjectures:

1) It can be shown that the two conjectures are equivalent.
2) The conjectures are natural generalizations of the results for complementary tree structures.
3) The "if" statements in the conjectures certainly are correct. (This follows from a slight modification of the proof of Theorem 9 and its dual.)

It is hoped that future research will either confirm these conjectures or else replace them with correct criteria for goodness.

VI. SOME IMPLICATIONS OF THE RESULTS

The results of the present paper are not only interesting in themselves, but also are of interest for the insight which they provide into the analysis and synthesis of RL networks.

For example, the reader will recall that the complex natural frequencies of any structure specify the poles of its $\left\{\begin{matrix}\text{pliers type admittances}\\ \text{soldering iron type impedances}\end{matrix}\right\}$.[6] Thus the fact that some RL structures possess independent natural frequencies, while others do not, implies that the same situation obtains for the poles of the $\left\{\begin{matrix}\text{pliers type admittances}\\ \text{soldering iron type impedances}\end{matrix}\right\}$ of such structures. Specifically the following conclusions are in order.

I. Barring cancellation, the poles of all $\left\{\begin{matrix}\text{pliers type admittances}\\ \text{soldering iron type impedances}\end{matrix}\right\}$ of reducible complementary tree structures are independent. That is, such poles can occur at virtually all sets of distinct points on the negative real axis of the s-plane.

II. Barring cancellation, the poles of all $\left\{\begin{array}{l}\text{pliers type admittances}\\ \text{soldering iron type impedances}\end{array}\right\}$ of non-reducible complementary tree structures are not independent. These poles are restricted not only to lie upon the negative real axis of the s-plane, but also are restricted by the *topological* condition (3).

Conclusions I and II have clear implications for network synthesis procedures that *do not involve cancellation*. These are as follows.

III. Networks derived from reducible complementary tree structures[7] quite likely have lenient realizability conditions. Thus such networks merit study for possible use in network synthesis procedures.

IV. Networks derived from non-reducible complementary tree structures[7] have stringent realizability conditions. Thus such networks should not be considered for use in network synthesis procedures.

It can reasonably be expected that future research will either confirm the conjectures of Section V or else replace them with correct criteria for goodness. Once this has been done, it should be possible to make statements analogous to I through IV for *all* RL networks.

VII. RC AND LC STRUCTURES

In conclusion, we should like to point out that all results of the present paper apply equally well to RC and LC structures.

To adapt the foregoing definitions, theorems and proofs to the RC and LC cases, it is only necessary to interchange concepts as follows.

RC Case:

resistance → susceptance
inductance → resistance

LC Case:

resistance → susceptance
inductance → inductance

$$-\sigma_i \longrightarrow \omega_i^2$$

APPENDIX

To enumerate the complementary tree structures of complexity n, it is only necessary to develop

a) a procedure for enumerating all distinct trees having n branches

and b) a procedure for enumerating all complementary tree structures which contain a given resistive tree and a given inductive tree.

A procedure of type a) makes it possible to construct all n branch resistive trees and all n branch inductive trees. A procedure of type b) makes it possible to construct all complementary tree structures appropriate to each pair of these trees.

An effective procedure of type a) is made available by the so-called tips. Every tree has at least one node, which is connected to the remainder of the tree by a single branch; the single branch which provides this connection is known as a tip. One property of tips is the following: If a tip is erased from an x branch tree, an x-1 branch tree results. It is immediately clear from this property that if one constructs all distinct x branch trees which can be formed by adding tips to x-1 branch trees, then one obtains all x branch trees. Thus to construct the trees required by a), one need only begin with a one-branch tree and repeat this procedure n-1 times.

To enumerate the complementary tree structures required by b), one can simply join together the given resistive tree and the given inductive tree in all possible ways. A systematic method for accomplishing this result is provided by the following steps.

i) Number the nodes of the resistive tree and of the inductive tree 1, 2, ... n+1.

ii) Pair off the integers 1, 2, ... n+1 against themselves in all possible ways.

iii) Interpret each pairing of integers in ii) as an instruction for joining nodes of the two trees, and then construct a complementary tree in accordance with the instruction.

FOOTNOTES

* This work was supported in part by the U.S. Army, Navy and Air Force under Contract DA 36-039-AMC-03200(E).

+ Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

1. F. M. Reza, "Order of Complexity and Minimal Structures in Network Analysis," U. of Illinois Symposium on Circuit Analysis, 7.1 - 7.3, 1955.

2. The reader might like to construct some complementary tree structures for himself. A simple procedure for doing this is given in the Appendix.

3. See Figures 2 and 4 for examples of this type of complementary tree structure.

4. Node splitting must be allowed throughout this procedure.

5. To guarantee success on the first trial, one can select the impedance levels as follows. Let ϵ denote the desired tolerance on the natural frequencies; let S_i denote the structure that results after the i^{th} resistor-inductor pair is inserted; let $z_i(s)$ denote the impedance of the i^{th} resistor-inductor pair, and let $Z_i(s)$ denote the impedance into which $z_i(s)$ looks in the structure S_i. Observe that the natural frequencies of S_i are given by the zeros of $Z_i(s) + z_i(s)$ or, equivalently, by the solutions of $Z_i(\sigma) = -z_i(\sigma)$.

   First, select the impedance level of the initial resistor-inductor pair at random. Next select the impedance level of the second resistor-inductor pair such that S_2 realizes the natural frequencies $s = -\sigma_1, -\sigma_2$ within the tolerance $\epsilon/(n - 1)$ (This can conveniently be done by referring to Figure 11; one determines the appropriate impedance level for $z_2(s)$ simply by scaling the dashed $-z_2(\sigma)$ curve down until it intersects the solid $Z_2(\sigma)$ curve at frequencies which differ from the values $s = -\sigma_1, -\sigma_2$ by amounts less than $\epsilon/(n - 1)$; that such scaling is possible follows from the fact that the zeros of $Z_2(s) + z_2(s)$ approach $s = -\sigma_1, -\sigma_2$ as the impedance level of $z_2(s)$ is reduced to zero). Next, select the impedance level of the third resistor-inductor pair such that S_3 realizes 1) the natural frequencies of S_2 and 2) the frequency $s = -\sigma_3$ within the tolerance $\epsilon/(n - 1)$. (Once again, this is conveniently done by referring to a figure analogous to Figure 11, in which both $Z_3(\sigma)$ vs. σ and $-z_3(\sigma)$ vs. σ are plotted.) Etc. At each state of this (n - 1 stage) process, one introduces a natural frequency within $\epsilon/(n - 1)$ of a desired natural frequency at the price of shifting the previously introduced natural frequencies by an amount not exceeding $\epsilon/(n - 1)$. Thus the natural frequencies of the final structure S_n differ from the desired natural frequencies $s = -\sigma_1, -\sigma_2, \ldots, -\sigma_n$ by amounts not exceeding $(n - 1)(\frac{\epsilon}{n - 1}) = \epsilon$.

6. E. A. Guillemin, Introductory Circuit Theory, New York, John Wiley & Sons, 1953, Pp. 430-431.

7. By means of soldering iron type and pliers type entries.

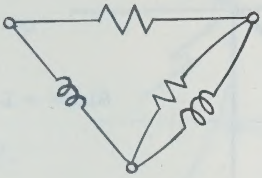


Figure 1.

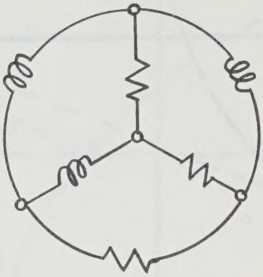


Figure 2.

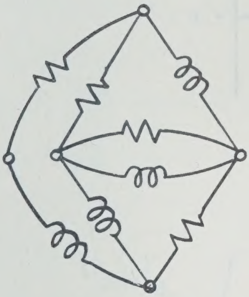


Figure 3.

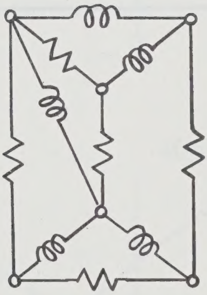


Figure 4.

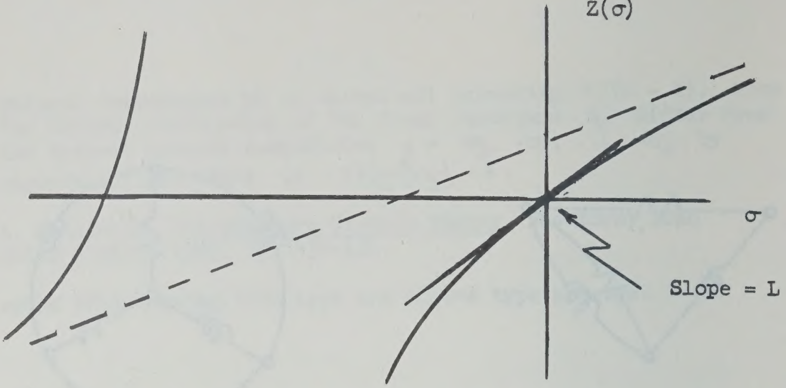


Figure 5.

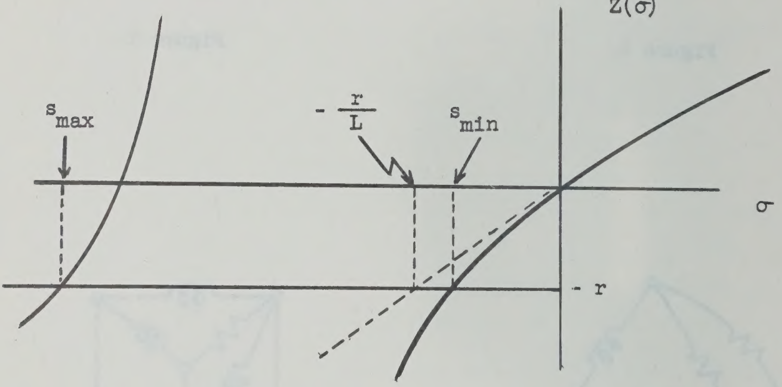


Figure 6.

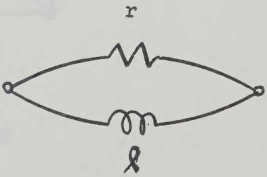


Figure 7.

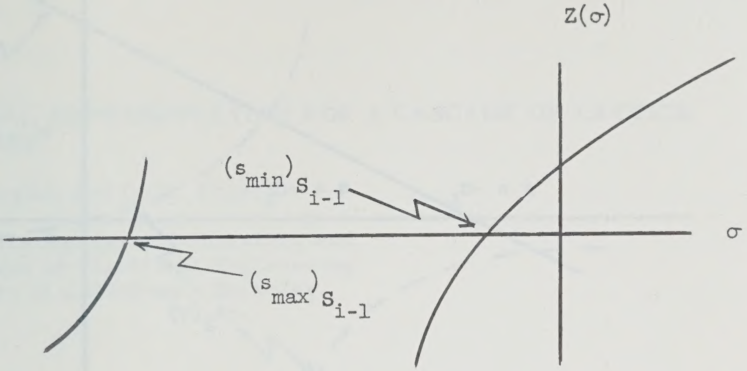

Z(σ)
$(s_{min})_{S_{i-1}}$
σ
$(s_{max})_{S_{i-1}}$

Figure 8.

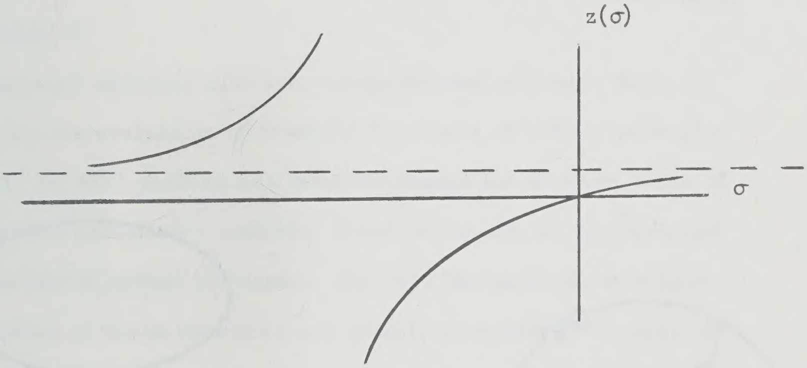


Figure 9.

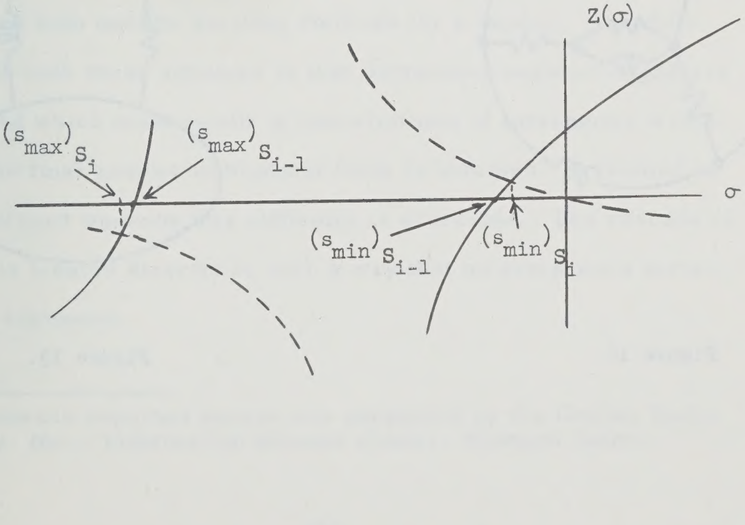

Z(σ)
$(s_{max})_{S_i}$
$(s_{max})_{S_{i-1}}$
σ
$(s_{min})_{S_{i-1}}$
$(s_{min})_{S_i}$

Figure 10.

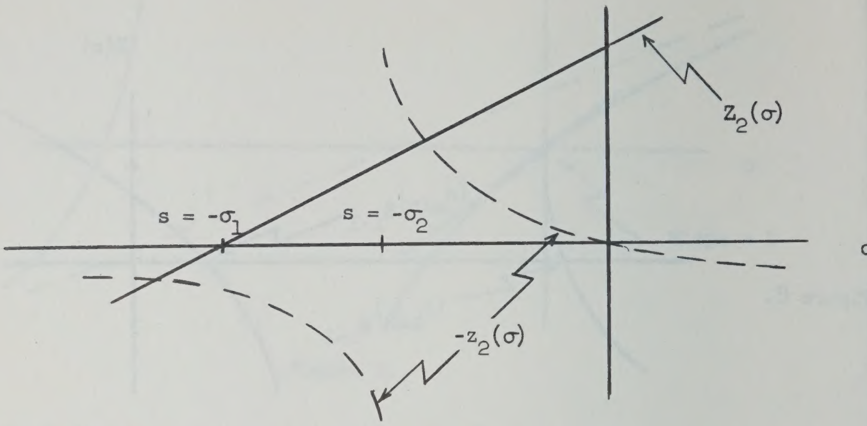


Figure 11.

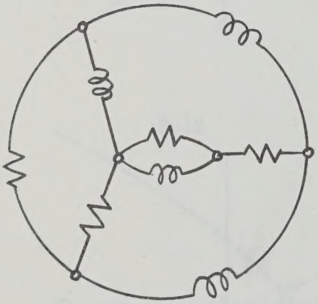


Figure 12.

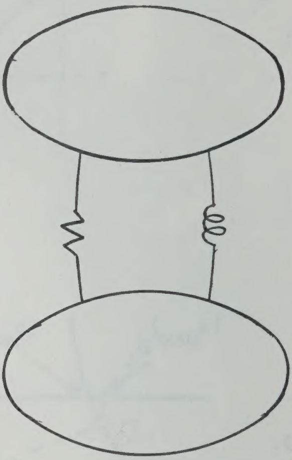


Figure 13.

A MINIMAL REPRESENTATION FOR A CASCADE OF LATTICE NETWORKS*

D. M. Layton and I. T. Frisch

Electronics Research Laboratory and
Department of Electrical Engineering
University of California, Berkeley

I. ABSTRACT

A number of highly efficient computational schemes have appeared for the evaluation of transfer functions of ladder networks. However, no such method has been developed for another class of highly useful networks - namely, those which can be represented as a cascade of lattice networks. Rather, the methods that have been applied to these networks are simply to multiply the cascade matrices of the individual lattices or to convert the network to a ladder and then use the existing methods for a ladder. The difficulty with both these schemes is that extraneous negative signs are introduced which must result in cancellations of extraneous terms before the final answer in minimal form is obtained. A method is now developed whereby this difficulty is overcome. The cascade of lattices is treated directly in such a way that no extraneous terms must be evaluated.

---

*The research reported herein was supported by the Collins Radio Company, Inc., Information Science Center, Newport Beach, Calif.

II. FORMULATION OF THE PROBLEM

Consider a network N_1 consisting of a cascade of lattices terminated in impedances Z_I and Z_L respectively as in Fig. 1.

Let $[Z]$ be the impedance matrix of N with row 1 and column 1 corresponding to $E_1(s)$ and $I_1(s)$ respectively and row 2 and column 2 corresponding to $E_2(s)$ and $I_2(s)$ respectively. Then we have

$$\frac{I_1(s)}{E_1(s)} = \frac{\Delta_{11}}{Z_I\Delta_{11}+\Delta} \tag{1}$$

and

$$\frac{I_2(s)}{E_1(s)} = \frac{\Delta_{12}}{Z_I\Delta_{11}+\Delta} \tag{2}$$

where Δ is the determinant of $[Z]$ and Δ_{ij} is the cofactor in the (i, j) position.

If $Z_I = 0$ then (1) and (2) reduce to (3) and (4) respectively.

$$\frac{I_1(s)}{E_1(s)} = \frac{\Delta_{11}}{\Delta} \tag{3}$$

$$\frac{I_2(s)}{E_1(s)} = \frac{\Delta_{12}}{\Delta} \ . \tag{4}$$

The problem then reduces to evaluating (1) – (4) in a minimal fashion.

III. EVALUATION OF Δ AND Δ_{11}

Wang[1] and Ting[2] have shown that the determinant of the loop impedance matrix of a network N consisting of resistors, inductors and capacitors, with e elements and v nodes can be evaluated by the following three steps:

1. Find a set of $(e - v + 1)$ independent loops.
2. Find the sum G_i of the impedances in the loop ℓ_i for $i = 1, \ldots (e - v + 1)$.
3. Δ is given by

$$\Delta = \prod_{i=1}^{(e - v + 1)} G_i \tag{5}$$

which is evaluated by using rules 1 and 2 below

$$\text{Rule 1:} \qquad Z_i + Z_j = \begin{cases} \text{the algebraic sum of } Z_i \text{ and } Z_j & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} .$$

$$\text{Rule 2:} \qquad Z_i \cdot Z_j = \begin{cases} \text{the algebraic product of } Z_i \text{ and } Z_j & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} .$$

We now apply these three steps to the network N in Figure 2 and make use of the symmetries and special properties of the network in order to find a minimal representation for Δ.

A set of $(e - v + 1) = (2n + 1)$ independent loops are chosen by the summations given in (6) – (10) where n is the number of lattices in N.

$$G_1 = Z_L + Z_{a_n} + Z_{b_n} \tag{6}$$

$$G_2 = Z_L + Z_{c_n} + Z_{d_n} \tag{7}$$

$$G_i = Z_{a_{n-k+1}} + Z_{d_{n-k+1}} + Z_{c_{n-k}} + Z_{d_{n-k}} \qquad i = 2k+1, k = 1, 2, \ldots, n-1 \tag{8}$$

and

$$G_i = Z_{a_{n-k+1}} + Z_{b_{n-k+1}} + Z_{c_{n-k+2}} + Z_{b_{n-k+2}} \quad , i = 2k, k = 2, 3, \ldots n. \tag{9}$$

$$G_{2n+1} = Z_{b_1} + Z_{c_1} \quad . \tag{10}$$

The loops are shown in Figure 2. It is easily shown that the loops chosen are independent since the network can be redrawn in planar form such that the loops form a set of meshes.

We then have

$$\Delta = \prod_{i=1}^{2n+1} G_i = (Z_L + Z_{a_n} + Z_{b_n})(Z_L + Z_{c_n} + Z_{d_n})(Z_{a_n} + Z_{d_{n-1}} + Z_{c_{n-1}} + Z_{d_n})$$

$$(Z_{a_{n-1}} + Z_{b_n} + Z_{c_n} + Z_{b_{n-1}})(Z_{a_{n-1}} + Z_{d_{n-2}} + Z_{c_{n-2}} + Z_{d_{n-1}})$$

$$(Z_{a_{n-2}} + Z_{b_{n-1}} + Z_{c_{n-1}} + Z_{b_{n-2}}), \ldots, (Z_{b_1} + Z_{c_1}) \cdot \tag{11}$$

Using rules 1 and 2 we have

$$\prod_{i=1}^{2} G_i = P_{n+1}(Z_{a_n} + Z_{b_n} + Z_{c_n} + Z_{d_n}) + S_{n+1}(Z_{a_n} + Z_{b_n})(Z_{c_n} + Z_{d_n}) \tag{12}$$

where $P_{n+1} \equiv Z_L$ and $S_{n+1} \equiv 1$.

Similarly

$$\prod_{i=1}^{4} G_i = P_n(Z_{a_{n-1}} + Z_{b_{n-1}} + Z_{c_{n-1}} + Z_{d_{n-1}}) + S_n(Z_{c_{n-1}} + Z_{d_{n-1}}) \cdot$$

$$(Z_{a_{n-1}} + Z_{b_{n-1}}). \tag{13}$$

where

$$P_n = P_{n+1}(Z_{a_n} + Z_{d_n})(Z_{b_n} + Z_{c_n}) + S_{n+1}\Big\{Z_{a_n} Z_{b_n} (Z_{c_n} + Z_{d_n})$$

$$+ Z_{c_n} Z_{d_n} (Z_{a_n} + Z_{b_n})\Big\} \tag{14}$$

and

$$S_n = P_{n+1}(Z_{a_n} + Z_{b_n} + Z_{c_n} + Z_{d_n}) + S_{n+1}(Z_{a_n} + Z_{b_n})(Z_{c_n} + Z_{d_n}). \tag{15}$$

Generalizing the above procedure we can then prove Theorem 1 by induction using Definition 1.

Definition 1:

$$P_i \equiv P_{i+1}(Z_{a_i} + Z_{d_i})(Z_{b_i} + Z_{c_i}) + S_{i+1}\left\{ Z_{a_i} Z_{b_i}(Z_{c_i} + Z_{d_i}) + Z_{c_i} Z_{d_i} (Z_{a_i} + Z_{b_i}) \right\} \tag{16}$$

$$S_i \equiv P_{i+1}(Z_{a_i} + Z_{b_i} + Z_{c_i} + Z_{d_i}) + S_{i+1}(Z_{a_i} + Z_{b_i})(Z_{c_i} + Z_{d_i})$$

$$S_{n+1} = 1, \quad P_{n+1} = Z_L , \tag{17}$$

Theorem 1:

$$\Delta = P_2(Z_{a_1} + Z_{d_1})(Z_{b_1} + Z_{c_1}) + S_2\left\{ Z_{a_1} Z_{b_1}(Z_{c_1} + Z_{d_1}) + Z_{c_1} Z_{d_1}(Z_{a_1} + Z_{b_1}) \right\}. \tag{18}$$

Proof: From the previous discussion the theorem is true for $n = 2$, i.e.,

$$\Delta = \prod_{i=1}^{4} G_i .$$

Assume next that the theorem is true for $n = k$, i.e., assume product of the first $2(k + 1)$ factors is

$$P_{n-k+1}(Z_{a_{n-k}} + Z_{b_{n-k}} + Z_{c_{n-k}} + Z_{d_{n-k}}) + S_{n-k}(Z_{c_{n-k}} + Z_{d_{n-k}})$$
$$(Z_{a_{n-k}} + Z_{b_{n-k}}) ,$$

and multiply by

$$(Z_{a_{n-k}} + Z_{d_{n-k}} + Z_{c_{n-k-1}} + Z_{d_{n-k-1}})(Z_{a_{n-k-1}} + Z_{b_{n-k}} + Z_{c_{n-k}} + Z_{b_{n-k-1}}$$

to give

$$P_{n-k}(Z_{a_{n-k-1}} + Z_{b_{n-k-1}} + Z_{c_{n-k-1}} + Z_{d_{n-k-1}}) + S_{n-k}(Z_{c_{n-k-1}} + Z_{d_{n-k-1}})$$
$$(Z_{a_{n-k-1}} + Z_{b_{n-k-1}}) ,$$

where

$$P_{n-k} = P_{n-k+1}(Z_{a_{n-k}} + Z_{d_{n-k}})(Z_{b_{n-k}} + Z_{c_{n-k}})$$
$$+ S_{n-k+1}\Big\{Z_{a_{n-k}} Z_{b_{n-k}}(Z_{c_{n-k}} + Z_{d_{n-k}}) + Z_{c_{n-k}} Z_{d_{n-k}}$$
$$(Z_{a_{n-k}} + Z_{b_{n-k}})\Big\} ;$$

and,

$$S_{n-k} = P_{n-k+1}(Z_{a_{n-k}} + Z_{b_{n-k}} + Z_{c_{n-k}} + Z_{d_{n-k}})$$
$$+ S_{n-k+1}(Z_{c_{n-k}} + Z_{d_{n-k}})(Z_{a_{n-k}} + Z_{b_{n-k}}) .$$

Finally we note that in (18) no further terms are zero because of rules 1 and 2, since no impedance is present in three loops. Hence the theorem is proved.

Δ_{11} is precisely the determinant of the impedance matrix of N with the source removed. Thus it is formed in exactly the same way as Δ, with the factor $(Z_{b_1} + Z_{c_1})$ omitted. We therefore have

Corollary 1.1:

$$\Delta_{11} = P_2(Z_{a_1} + Z_{b_1} + Z_{c_1} + Z_{d_1}) + S_2(Z_{a_1} + Z_{b_1})(Z_{c_1} + Z_{d_1}). \qquad (19)$$

IV. EVALUATION OF Δ_{12}

Theorem 2:

$$\Delta_{12} = \prod_{i=1}^{2n+1} (Z_{a_i} Z_{c_i} - Z_{b_i} Z_{d_i}) .$$

Proof: We can readily show that for a single lattice

$$\Delta_{12} = (Z_{a_n} Z_{c_n} - Z_{b_n} Z_{d_n}) \,. \tag{20}$$

We next show the theorem is true for the $(k+1)$ stages in Figure 3 after assuming it true for k stages.

For the currents and voltages in Figure 3 the impedance mattix can be written as in (21), where the z'_{ij}'s are the entries in the loop matrix of the k stages and I'_i, E'_i, $i = 1, \ldots, 2k+1$ are the loop currents and voltages defining the z'_{ij}'s. Δ_{12} is then the determinant of the matrix formed by crossing out row 2 and column 1. Using a Laplace expansion about the first row of the matrix thus formed we then have Δ_{12} as in (22). Using a Laplace expansion about the first row of each determinant (22) we finally have

$$\Delta_{12} = \left[(Z_{a_n} + Z_{b_n})(Z_{a_n} + Z_{d_n}) - Z_{a_n}(Z_{a_n} + Z_{b_n} + Z_{c_n} + Z_{d_n}) \right] \Delta'_{12}$$

$$= (Z_{a_n} Z_{c_n} - Z_{b_n} Z_{d_n}) \Delta'_{12} \quad , \tag{23}$$

where Δ'_{12} is the minor in the 12 position of Δ', the determinant of the impedance matrix for k stages. This completes the proof.

V. SYMMETRIC LATTICES

For the special case in which all the lattices are symmetric we have $Z_{a_i} = Z_{c_i}$ and $Z_{b_i} = Z_{d_i}$, for $i = 1, \ldots, n$. (18), (19) and (20) reduce to (24), (25) and (26) respectively, where the bars refer to the symmetric case

$$\overline{\Delta} = \overline{P}_2 (Z_{a_1} + Z_{b_1})^2 + 2\overline{S}_2 Z_{a_1} Z_{b_1} (Z_{a_1} + Z_{b_1}) \tag{24}$$

$$\left[\begin{array}{c} E_1 \\ E_2 \\ E_3 \\ E_4 \\ \hline E_5 \\ \vdots \\ E_{2k+3} \end{array}\right] = \left[\begin{array}{cccc|ccc} z'_{11} & 0 & 0 & z'_{12} & z'_{13} & \cdots & z'_{1,\,2k+1} \\ 0 & (Z_{a_n} + Z_{b_n}) & (Z_{a_n} + Z_{b_n}) & -Z_{a_n} & 0 & \cdots & 0 \\ 0 & (Z_{a_n} + Z_{b_n}) & \begin{pmatrix} Z_{a_n} + Z_{b_n} \\ + Z_{c_n} + Z_{d_n} \end{pmatrix} & -(Z_{a_n} + Z_{d_n}) & 0 & \cdots & 0 \\ z'_{21} & -Z_{a_n} & -(Z_{a_n} + Z_{d_n}) & (Z_{a_n} + Z_{d_n} + z'_{22}) & z'_{23} & \cdots & z'_{2,\,2k+1} \\ \hline z'_{31} & 0 & 0 & z'_{32} & z'_{33} & \cdots & z'_{3,\,2k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ z'_{2k+1,1} & 0 & 0 & z'_{2k+1,\,2} & z'_{2k+1,\,3} & \cdots & z'_{2k+1,\,2k+1} \end{array}\right] \left[\begin{array}{c} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ \vdots \\ I_{2k+3} \end{array}\right] \tag{21}$$

$$\Delta_{12} = -(Z_{a_n} + Z_{b_n}) \left| \begin{array}{cc|ccc} 0 & -(Z_{a_n} + Z_{d_n}) & 0 & \cdots & 0 \\ z'_{21} & +\begin{pmatrix} Z_{a_n} + Z_{d_n} \\ +z'_{22} \end{pmatrix} & z'_{23} & \cdots & z'_{2,\,2k+1} \\ \hline z'_{31} & z'_{32} & z'_{33} & \cdots & z'_{3,\,2k+1} \\ \vdots & \vdots & \vdots & & \vdots \\ z'_{2k+1,\,1} & z'_{2k+1,\,2} & z'_{2k+1,\,3} & \cdots & z'_{2k+1,\,2k+1} \end{array} \right|$$

$$-Z_{a_n} \left| \begin{array}{cc|ccc} 0 & \begin{pmatrix} Z_{a_n} + Z_{b_n} \\ +Z_{c_n} + Z_{d_n} \end{pmatrix} & 0 & \cdots & 0 \\ z'_{21} & -(Z_{a_n} + Z_{d_n}) & z'_{23} & \cdots & z'_{2,\,2k+1} \\ \hline z'_{31} & 0 & z'_{33} & \cdots & z'_{3,\,2k+1} \\ \vdots & \vdots & \vdots & & \vdots \\ z'_{2k+1,\,1} & 0 & z'_{2k+1,\,3} & \cdots & z'_{2k+1,\,2k+1} \end{array} \right| \qquad (22)$$

$$\bar{\Delta}_{11} = 2\,\bar{P}_2(Z_{a_1} + Z_{b_1}) + \bar{S}_2(Z_{a_1} + Z_{b_1})^2 \tag{25}$$

$$\bar{\Delta}_{12} = \prod_{i=1}^{n} (Z_{a_i}^2 - Z_{b_i}^2) \tag{26}$$

where

$$\bar{P}_i = \bar{P}_{i+1}(Z_{a_i} + Z_{b_i})^2 + 2\bar{S}_{i+1} \cdot Z_{a_i} Z_{b_i}(Z_{a_i} + Z_{b_i}) \tag{27}$$

and

$$\bar{S}_i = 2\,\bar{P}_{i+1}(Z_{a_i} + Z_{b_i}) + \bar{S}_{i+1}(Z_{a_i} + Z_{b_i})^2 \tag{28}$$

$$\bar{P}_{n+1} = Z_L$$

$$\bar{S}_{n+1} = 1.$$

A flow chart for calculating driving point and transfer functions is given for the unsymmetric and the symmetric cases in Table I.

VI. MINIMALITY OF EXPANSIONS

It is apparent from (18) and (19) that cancellations cannot occur by subtraction in Δ_{11} and Δ since all terms appear only with plus signs. Furthermore from (20) no cancellations can occur in Δ_{12} by subtraction since each factor in the product consists of terms with different subscripts and no cancellations can occur within a given factor. The next question is whether Δ, Δ_{11}, and Δ_{12} may have common factors which would cause cancellations by division. The factors of Δ_{12} are of the form $(Z_{a_i} Z_{c_i} - Z_{b_i} Z_{d_i})$ and these are not factors of Δ_{11} or Δ. Thus the only question that remains is whether Δ_{11} and Δ have common factors. They clearly do not, since they would have to be contained in both the P_i and S_i, $i = 2, \ldots, n$, and an examination of these expressions indicate no common factors.

For the case of symmetric lattices the same argument shows that there is no cancellation by subtraction in (24), (25), and (26). On the other hand Δ, Δ_{11}, and Δ_{12} all contain the common factor.

$$\prod_{i=1}^{n} (Z_{a_i} + Z_{b_i}) = (Z_{a_1} + Z_{b_1})(Z_{a_2} + Z_{b_2}) \dots (Z_{a_n} + Z_{b_n}).$$

Thus in performing calculations for symmetric lattices $\bar{\Delta}$, $\bar{\Delta}_{11}$, and $\bar{\Delta}_{12}$ can be replaced by Δ^*, Δ_{11}^* and Δ_{12}^* respectively where

$$\Delta^* = \frac{\bar{\Delta}}{\prod_{i=1}^{n} (Z_{a_i} + Z_{b_i})} = P_2^*(Z_{a_1} + Z_{b_1}) + 2\, S_2^*\, Z_{a_1} Z_{b_1} \tag{29}$$

$$\Delta_{11}^* = \frac{\bar{\Delta}_{11}}{\prod_{i=1}^{n} (Z_{a_i} + Z_{b_i})} = 2\, P_2^* + S_2^*\, (Z_{a_1} + Z_{b_1}) \tag{30}$$

$$\Delta_{12}^* = \frac{\bar{\Delta}_{12}}{\prod_{i=1}^{n} (Z_{a_i} + Z_{b_i})} = \prod_{i=1}^{n} (Z_{a_i} - Z_{b_i}) \tag{31}$$

and

$$P_i^* = P_{i+1}^*\, (Z_{a_i} + Z_{b_i}) + 2\, S_{i+1}^*\, Z_{a_i} Z_{b_i} \tag{32}$$

$$S_i^* = 2\, P_{i+1}^* + S_{i+1}^*\, (Z_{a_i} + Z_{b_i}) \,, \tag{33}$$

$$P_{n+1}^* = Z_L, \quad S_{n+1}^* = 1.$$

The representations are then minimal.

VII. EXAMPLE

For the network in Figure 4 we have

$$\frac{E_1(s)}{I_1(s)} = \frac{Z_I\Delta_{11}^* + \Delta^*}{\Delta_{11}^*} \quad , \quad \frac{E_1(s)}{I_2(s)} = \frac{Z_I\Delta_{11}^* + \Delta^*}{\Delta_{12}^*} \tag{34}$$

Since the lattices are symmetric we use (29) - (33) with $n = 2$.

$$P_2^* = Z_L(Z_{a_2} + Z_{b_2}) + 2\,Z_{a_2}Z_{b_2}$$

$$S^* = 2\,Z_L + Z_{a_2} + Z_{b_2} \tag{35}$$

thus

$$\begin{aligned}\Delta^* &= P_2^*(Z_{a_1} + Z_{b_1}) + 2S_2^*\,Z_{a_1}\,Z_{b_1}\\ &= Z_L(Z_{a_2} + Z_{b_2})(Z_{a_1} + Z_{b_1}) + 2\,Z_{a_2}Z_{b_2}(Z_{a_1} + Z_{b_1}) + 4Z_L Z_{a_1} Z_{b_1}\\ &\quad + 2\,Z_{a_1}Z_{b_1}(Z_{a_2} + Z_{b_2})\\ &= Z_L Z_{a_2} Z_{a_1} + Z_L Z_{b_2} Z_{a_1} + Z_L Z_{a_2} Z_{b_1} + Z_L Z_{b_2} Z_{b_1} + 2Z_{a_1} Z_{a_2} Z_{b_2}\\ &\quad + 2Z_{a_2} Z_{b_2} Z_{b_1} + 4Z_L Z_{a_1} Z_{b_1} + 2Z_{a_1} Z_{b_1} Z_{a_2} + 2Z_{a_1} Z_{b_1} Z_{b_2}\end{aligned} \tag{36}$$

$$\begin{aligned}\Delta_{11}^* &= 2\,P_2^* + S_2^*\,(Z_{a_1} + Z_{b_1})\\ &= 2\,Z_L(Z_{a_2} + Z_{b_2}) + 4Z_{a_2} Z_{b_2} + (2Z_L + Z_{a_2} + Z_{b_2})(Z_{a_1} + Z_{b_1})\\ &= 2\,Z_L Z_{a_2} + 2Z_L Z_{b_2} + 4Z_{a_2} Z_{b_2} + 2Z_{a_1} Z_L + 2Z_{b_1} Z_L + Z_{a_1} Z_{a_2}\\ &\quad + Z_{b_1} Z_{a_2} + Z_{a_1} Z_{b_2} + Z_{b_1} Z_{b_2}\end{aligned} \tag{37}$$

$$\Delta_{12}^* = (Z_{a_2} - Z_{b_2})(Z_{a_1} - Z_{b_1}) = Z_{a_2} Z_{a_1} - Z_{b_2} Z_{a_1} - Z_{a_2} Z_{b_1} + Z_{b_2} Z_{b_1}.$$

VIII. REFERENCES

1. K. T. Wang, "On a New Method of Analysis of Electrical Networks," Memoir 2, Nat'l. Research Institute of Engineering, Academia Sinica (1934).
2. S. L. Ting, "On the General Properties of Electrical Network Determinants," Chinese Journal of Physics, Vol. 1 (1935).

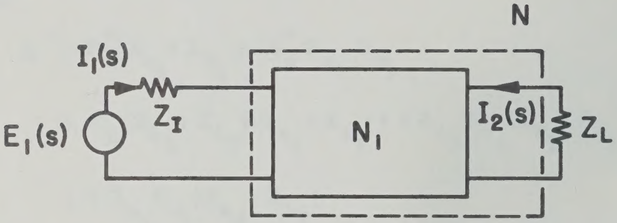

N
$I_1(s)$
Z_I
$E_1(s)$
N_1
$I_2(s)$
Z_L

Fig. 1. The model

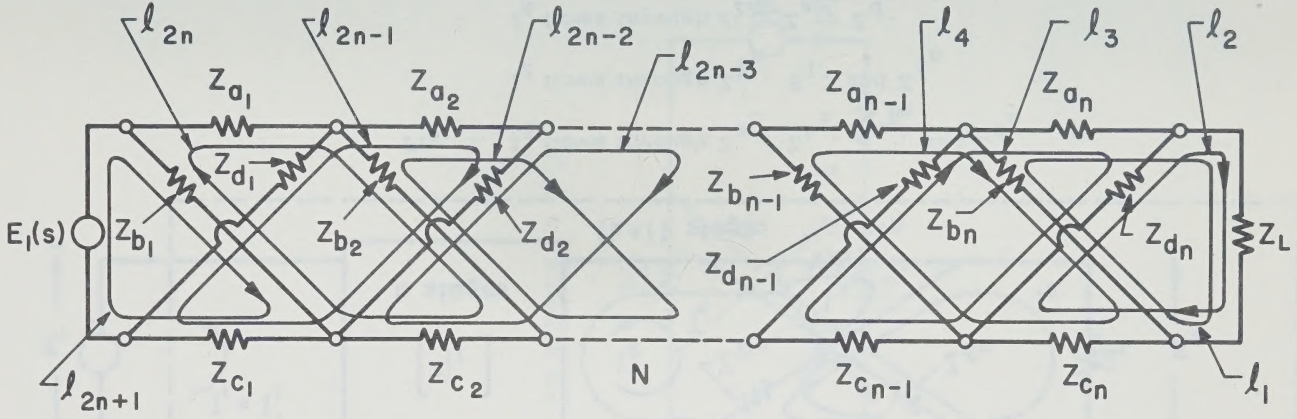

ℓ_{2n}
ℓ_{2n-1}
ℓ_{2n-2}
ℓ_{2n-3}
ℓ_4
ℓ_3
ℓ_2
Z_{a_1}
Z_{a_2}
$Z_{a_{n-1}}$
Z_{a_n}
Z_{d_1}
$Z_{b_{n-1}}$
$E_1(s)$
Z_{b_1}
Z_{b_2}
Z_{d_2}
Z_{b_n}
Z_{d_n}
Z_L
$Z_{d_{n-1}}$
Z_{c_1}
Z_{c_2}
N
$Z_{c_{n-1}}$
Z_{c_n}
ℓ_{2n+1}
ℓ_1

Fig. 2. Definition of loops

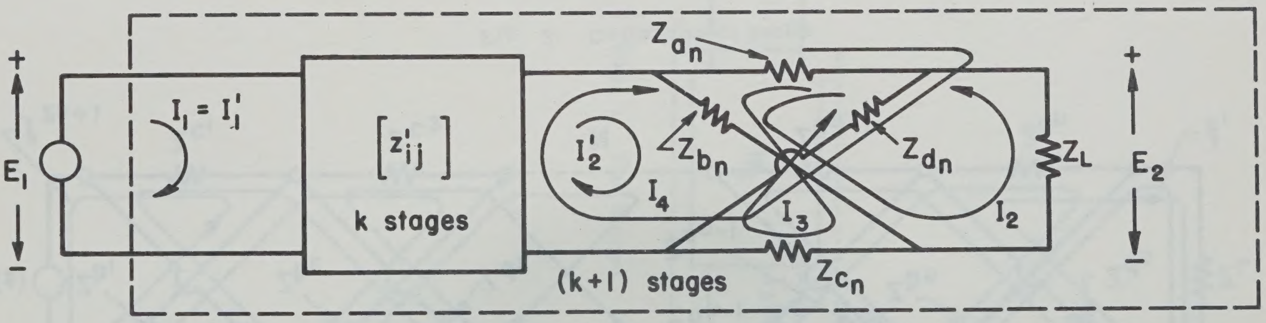
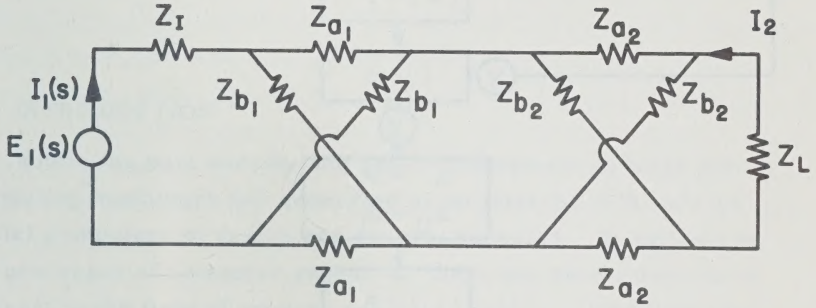

+
E_1
−
$I_1 = I_1'$
$[z_{ij}']$
k stages
I_2'
I_4
Z_{a_n}
Z_{b_n}
I_3
Z_{c_n}
Z_{d_n}
I_2
Z_L
(k+1) stages
+
E_2
−

Fig. 3. I_3 flows through Z_{a_n}, Z_{b_n}, Z_{c_n}, and Z_{d_n}

I_2 flows through Z_{b_n}, Z_L, and Z_{a_n}

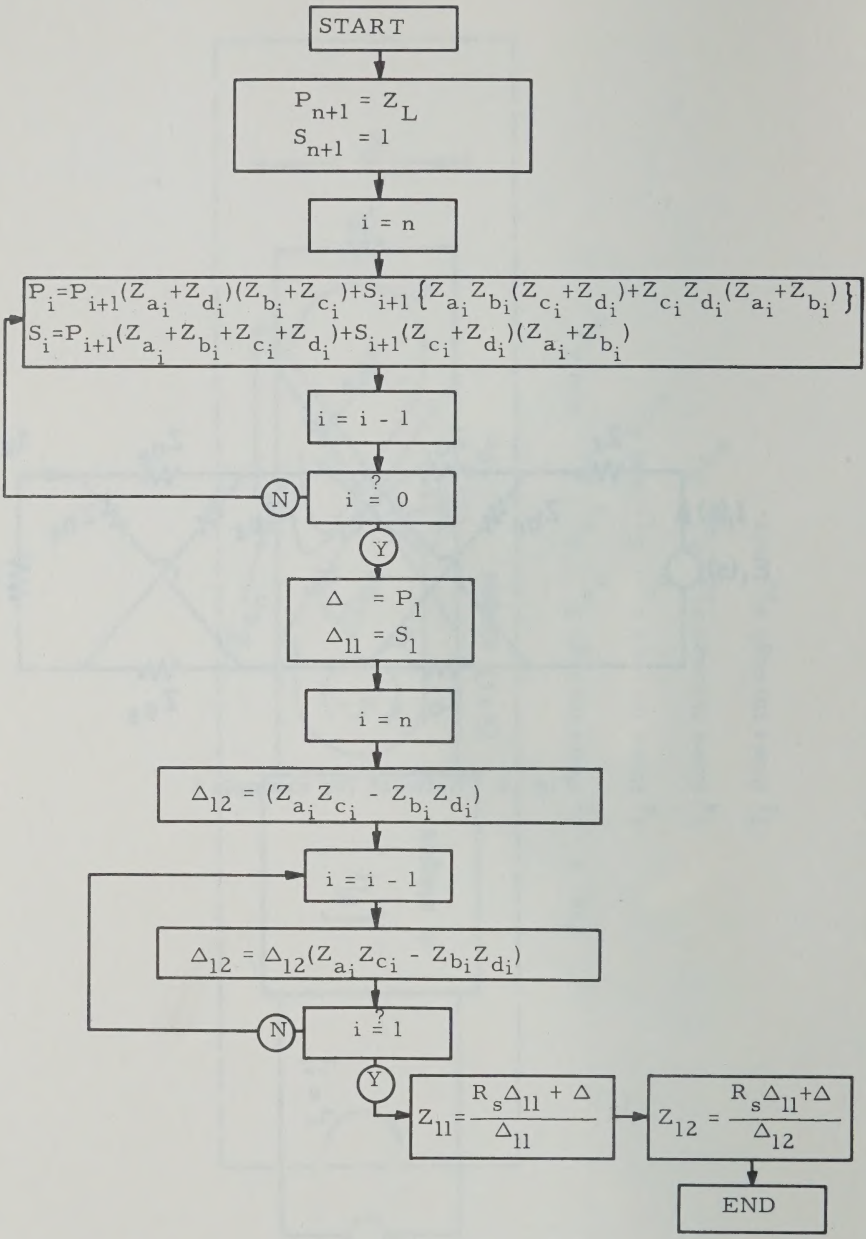
I_4 flows through z_{22}', Z_{a_n}, Z_{d_n}

I_2' flows through z_{22}' only.



Z_I
Z_{a_1}
Z_{a_2}
I_2
$I_1(s)$
$E_1(s)$
Z_{b_1}
Z_{b_1}
Z_{b_2}
Z_{b_2}
Z_L
Z_{a_1}
Z_{a_2}

Fig. 4. Network for example 1


START
$P_{n+1} = Z_L$
$S_{n+1} = 1$
$i = n$
$P_i = P_{i+1}(Z_{a_i}+Z_{d_i})(Z_{b_i}+Z_{c_i})+S_{i+1}\{Z_{a_i}Z_{b_i}(Z_{c_i}+Z_{d_i})+Z_{c_i}Z_{d_i}(Z_{a_i}+Z_{b_i})\}$
$S_i = P_{i+1}(Z_{a_i}+Z_{b_i}+Z_{c_i}+Z_{d_i})+S_{i+1}(Z_{c_i}+Z_{d_i})(Z_{a_i}+Z_{b_i})$
$i = i - 1$
$i \overset{?}{=} 0$
N
Y
$\Delta = P_1$
$\Delta_{11} = S_1$
$i = n$
$\Delta_{12} = (Z_{a_i}Z_{c_i} - Z_{b_i}Z_{d_i})$
$i = i - 1$
$\Delta_{12} = \Delta_{12}(Z_{a_i}Z_{c_i} - Z_{b_i}Z_{d_i})$
$i \overset{?}{=} 1$
N
Y
$Z_{11} = \frac{R_s\Delta_{11} + \Delta}{\Delta_{11}}$
$Z_{12} = \frac{R_s\Delta_{11} + \Delta}{\Delta_{12}}$
END

Flow chart for unsymmetric lattices

ANALYSIS OF NETWORKS BY VERTEX REDUCTION*

Howard Frank
Department of Electrical Engineering
Northwestern University
Evanston, Illinois

I. INTRODUCTION

Within the past decade, the rapid development of large scale computing machinery has generated great interest in the use of digital computers to design and analyze networks. In addition to the new areas of computer research, there has been a revival of interest in the field of network analysis [1-9] . Traditionally, simultaneous loop or node equations were used to analyze networks. The necessity of writing simultaneous equations was eliminated by the application of the graph theoretic properties of general networks. Topological formulas have been developed which relate a network's driving point and transfer characteristics to the structure of the network. These network functions may then be expressed as ratios of tree, two-tree, and three-tree admittance products [1] . Thus, topological network analysis reduces to the problem of generating the trees, two trees and three trees of a graph [2, 3] .

---

*This work was supported by the U.S. Air Force Office of scientific Research, Grant No. AF-AFOSR-98-63.

Efficient tree generation has been the subject of a good deal of recent research in network analysis [4-7] . As yet present techniques of tree generation are not entirely satisfactory from the standpoint of computer storage time and programming simplicity. This situation is further complicated in the case of active networks, where two graphs are required for writing topological formulas and factors in the formulas may have negative signs [8, 9] . Although the use of topological formulas has extended the range of analyzable networks, there is still a great deal to be accomplished in the way of fast, easily programmed techniques to analyze networks. Consequently, a re-evaluation of the traditional tools of network analysis is desirable.

It is well known that a purely resistive network may be readily reduced to a single resistive loop through the use of generalized star mesh transformations. The relative simplicity of the star mesh transformation gives reason to believe that its use should be valuable in reducing the order of complexity of networks not restricted to the resistive case.

This paper will show that the star mesh transformation, in conjuction with such elementary techniques as source transformations and the "pliars entry" [10] , can be utilized to develop a method to reduce any lumped, linear network to an equivalent one loop representation. In addition, the applicability of the method to digital computers is explored by indicating several interesting theoretical problems.

II. PRELIMINARY CONSIDERATIONS

A general passive RLC network without mutual inductances consists of an arbitrary number of passive elements and independent voltage and current sources. It is meaningless to consider networks which have voltage sources of different values in parallel or current sources of different values in series. If a voltage sources is in parallel with any other type branch the entire configuration may be replaced by that voltage source. Similarly, a current source in series with any other elements is equivalent to the current source alone.

For an element to have any effect in the network, the degree of both of its vertices must be greater than one. All passive branches in series or parallel may be combined and represented by single impedances. Any parallel combinations of active and passive elements may also be combined into a series combination of a passive element and an independent voltage source (the Thevenin equivalent). Therefore, without loss of generality, any passive network can be replaced by an equivalent one with no parallel branches.

In order that unique solutions exist for the various network currents and voltages, there must be a tree of the network such that the voltage sources are contained in the branches of the tree and the current sources are contained in the chord set of the tree [1]. Consequently, if a current driver I_{ij} is incident at vertices v_i and v_j, there must be a path in the network between these vertices, not including I_{ij}. This property can be used to replace I_{ij} by an open circuit and place current sources of the same value and orientation in parallel with all the branches in the path from v_i to v_j. The revised graph is equivalent to the original one. At this point, all the current sources can be transformed to voltage sources. In an identical fashion, voltage sources may be pushed through vertices to produce new equivalent graphs. These techniques are termed the "soldering iron" method and the "pliars entry" method [10] . By continued application of both source transformations and the soldering iron technique, all current sources may be removed from the graph. Hence, the general RLC network can be replaced by one with parallel branches or current sources.

One additional tool which we will need to introduce is the generalized star mesh transformation. Suppose n admittances $Y_1, Y_2, , , , , Y_n$ are connected in a star at terminal v_o. An n terminal mesh can be constructed such that every vertex is connected to every other vertex. If the admittance of the branch connecting v_k to v_m is Y_{km} given by (1), then equal currents will flow if the (n + 1) terminal star is replaced by the n terminal mesh. Hence, vertex v_o has been eliminated at the expense of forming a complete network of n(n - 1)/2 branches [11].

$$Y_{km} = \frac{Y_k Y_m}{Y_1 + Y_2 + \ldots + Y_n} \tag{1}$$

III. REDUCTION OF PASSIVE NETWORKS

In the previous section we have shown that any passive RLC network without mutuals may be represented by a graph with no current sources or parallel branches. Suppose it is desired to determine the voltage across two terminals of a network without solving a large number of simultaneous equations or without generating the trees of the graph. In some way, it is then necessary to simplify the network in an uncomplicated fashion.

If there are only passive branches incident at a vertex v_i in the network, the star mesh transformation may be immediately applied and v_i eliminated from the graph. Parallel branches generated by the transformation may be combined with the existing network. This process can be continued until a vertex is reached which has voltage source incident. When this situation occurs, (Figure 1) the star mesh transformation can not be directly applied.

Let v_j and v_k be the incident vertices of a voltage source E_{jk}. Suppose it is desirable to eliminate v_j. If either vertex has degree one, E_{jk} may be immediately discarded. If v_k has degree two, then the other branch incident at v_k is either active or passive. If active, the two voltage sources may be combined and the problem reduces to the original one. If passive, the position of the voltage source and the passive element may be interchanged and the problem is eliminated. Thus, unless the degree of v_k is at least three, there are only trivial restrictions on the application of a star mesh transformation.

Consider the case where the degree of v_k is at least three. Using the pliars entry technique, E_{jk} can be pushed through v_k and then v_j and v_k shorted together and all parallel branches combined. If the degree of v_k was originally q, there are now q-1 voltage sources incident at the new vertex v_j'. Let the vertices at distance one from v'_j with these incident voltage sources be $v_{j1}, v_{j2}, \ldots, v_{jq-1}$ and those with purely passive branches be $v_{k1}, v_{k2}, \ldots, v_{kr}$. Clearly $v_{ji} \neq v_{km}$ for all i and m, since if $v_{jy} = v_{kz}$ then there are two parallel branches between v'_j and v_{jy}. Since all parallel branches have already been combined, this can not happen. Consequently, if the voltage sources are pushed through $v_{j1}, v_{j2}, \ldots, v_{jq-1}$, there are no longer any voltage sources incident at v'_j. V'_j may now be reduced. Figures 2 and 3 demonstrate the process.

Eliminating v_j generates a number of new passive branches, some of which may be in parallel with old elements. These must be combined. The process can then be continued until only three vertices remain, two of which are the original output vertices. A network of one loop has thus been obtained.

One problem remains before the output voltage can be determined. The original output terminals could have been "obscured" by successive application of the reduction technique. If the original network had a load, or a load developed during the process, voltage sources might have been pushed through the output terminals, changing the voltage between the terminals. To compensate for the voltages pushed through the loaded output vertices, it is only necessary to sum these voltages and adjust the output voltage by this amount.

In the general case, the output voltage may be determined by solving for the voltage between vertices v_n and v_{n-1} of Figure 4 and then subtracting the sum of the voltages pushed through the output vertices.

$$V_{out} = \frac{E_{n-2,n-1} + E_{n-1,n} - E_{n-2,n}}{Z_{n-2,n} + Z_{n-2,n-1} + Z_{n-1,n}} Z_{n-1,n} - E_{n-1,n} - V_{n-1,n} \tag{2}$$

where $V_{n-1,n}$ is the difference of the sum of the voltages pushed through vertices v_n and v_{n-1}, and $E_{i,j}$ and $Z_{i,j}$ are the various voltages and impedances of the final one loop network.

$$\text{If } y_{n-1,n} = 0, \text{ then } V_{out} = E_{n-2,n-1} - E_{n-2,n} \tag{3}$$

The above technique can be described by a systematic matrix procedure. In general, a branch of any network may be considered to consist of two series elements, a passive admittance Y and an independent voltage driving point source E. Thus, a typical branch b_{ij}, incident at vertices v_i and v_j is completely specified by the ordered pair, $b_{ij} = (Y_{ij}, E_{ij})$. ($E_{ij}$ is defined to be positive if oriented towards v_j.)

Using this representation, the following conventions will be adopted:

1. If there is no connection between v_i and v_j, $b_{ij} = (0, 0)$
2. If there is only a passive branch between v_i and v_j, $b_{ij} = (Y_{ij}, 0)$
3. If there is only a voltage source between v_i and v_j, $b_{ij} = (0, E_{ij})$
4. If v_i and v_j are shorted, $b_{ij} = (X, 0)$

With the above terminology, the connection matrix C is defined such that

$$\begin{aligned} c_{ij} &= b_{ij} = (Y_{ij}, E_{ij}) && i < j \\ c_{ij} &= (X, 0) && i = j \\ c_{ij} &= (Y_{ji}, -E_{ji}) && i > j \end{aligned} \tag{4}$$

Obviously, C completely describes its corresponding graph, and several matrix operations may now be defined.

A. If $c_{ij} = (Y_{ij}, E_{ij})$, an equivalent matrix can be found corresponding to the network obtained by pushing E_{ij} through v_j. In the matrix, replace c_{ij} by c_{ij}' and $c_{jk} = (Y_{jk}, E_{jk})$ by c_{jk}', where

$$c_{ij}' = (Y_{ij}, 0)$$

and, for all $k \neq j$ such that $c_{jk} \neq (0, 0)$, $c_{jk} \neq (X, 0)$ (5)

$$\begin{aligned} c_{jk}' &= (Y_{jk}, E_{jk} + E_{ij}) && k > j \\ c_{jk}' &= (Y_{jk}, E_{ij} - E_{jk}) && k < j \end{aligned}$$

B. If $c_{ij} = (0, E_{ij})$, c_{jk} may be replaced by c_{jk}' as in A and then v_i is shorted to v_j by adding term by term the i-th row of the matrix to the j-th row and then removing the i-th rows and columns. Addition is defined as

$$c_{jr}^{+} = c_{ir} + c_{jr} = \left(Y_{ir} + Y_{jr}, \frac{Y_{ir}E_{ir} + Y_{jr}E_{jr}}{Y_{ir} + Y_{jr}}\right) \tag{6}$$

for $c_{ir}, c_{jr} \neq (0, E), (X, 0)$

$$\left.\begin{aligned} (0, E_1) + (Y, E_2) &= (0, E_1) \\ (X, 0) + (Y, E) &= (X, 0) \end{aligned}\right\} \quad Y \neq 0$$

For $Y = 0$, the solution is indeterminant unless $E_1 = E_2$.

C. A star mesh transformation can be performed directly on the connection matrix. A row i in the matrix such that $E_{ij} = 0$ for all j, is equivalent to a vertex v_i with only passive branches incident

If these branches have admittances $Y_{i,j1}, Y_{i,j2}, \ldots, Y_{i,jp}$, the new entries of the connection matrix are found by replacing $c_{jq,jr}$ by $c^*_{jq,jr}$ where

$$c^*_{jq,jr} = \left(Y_{jq,jr} + \frac{Y_{i,jq} Y_{i,jr}}{\sum_{j=1}^{n} Y_{ij}} \; , \; \frac{Y_{jq,jr} E_{jq,jr}}{Y_{jq,jr} + \frac{Y_{i,jq} Y_{i,jr}}{\sum_{j=1}^{n} Y_{ij}}} \right) \qquad (7)$$

The $c^*_{jq,jr}$ are determined by taking all combinations of the jk's two at a time. After all the $c^*_{jq,jr}$'s are formed, the i-th rows and columns are removed from the matrix.

D. Whenever operations A or B are applied, sources may be pushed through vertices v_{n-1} or v_n (the output vertices). Only the voltage sources $E_{i,n-1}$ or $E_{i,n}$ are involved. Therefore, the voltage $V_{n-1,n}$ is adjusted every time a source is pushed through an output terminal, where $V_{n-1,n}$ is defined by (8).

$$V_{n-1,n} = 0 + E_{i,n} - E_{i,n-1} \qquad (8)$$

By using the above operations a systematic procedure can be developed for reducing the matrix. The vertices of the graph $v_1, v_2, \ldots, v_n$ are labeled arbitrarily, except for v_{n-1} and v_n which are the specified output terminals. First, row 1 of the connection matrix is operated on with B. Then A is applied repeatedly until $E_{ij} = 0$ for all j. For $j = n-1$ and $j = n$, D is performed. Then C is applied. The matrix has now been reduced to $(n-1) \times (n-1)$. Row 2 is operated on next and the process is continued until a 3x3 matrix is left. The desired output voltage is then calculated from Equation (2) or (3). The following example illustrates one complete step of the procedure.

The network to be analyzed is shown in Figure 5 and its associated matrix is given below in (a). The quantity to be determined is the voltage V_{out} between vertices v_4 and v_5.

$$C=\begin{bmatrix} (X,0) & (y_1,e_1) & (y_2,0) & (y_3,e_3) & (y_4,0) \\ (y_1,-e_1) & (X,0) & (y_5,e_5) & (0,0) & (y_6,0) \\ (y_2,0) & (y_5,-e_5) & (X,0) & (y_7,0) & (y_i,e_8) \\ (y_3,-e_3) & (0,0) & (y_7,0) & (X,0) & (y_{10},0) \\ (y_4,0) & (y_7,0) & (y_8,-e_8) & (y_{10},0) & (X,0) \end{bmatrix}$$

(a)

$$C=\begin{bmatrix} (X,0) & (y_1,0) & (y_2,0) & (y_3,0) & (y_4,0) \\ (y_1,0) & (X,0) & (y_5,e_5+e_1) & (0,0) & (y_6,e_1) \\ (y_2,0) & (y_5,-e_5-e_1) & (X,0) & (y_7,e_3) & (y_8,e_8+e_3) \\ (y_3,0) & (0,0) & (y_7,-e_3) & (X,0) & (y_{10},0) \\ (y_4,0) & (y_7,-e_3) & (y_8,-e_8-e_3) & (y_{10},0) & (X,0) \end{bmatrix}$$

(b)

$$C=\begin{bmatrix} (X,0) & (Y_1,\frac{(e_5+e_1)y_5}{Y_1}) & (Y_2,0) & (Y_3,\frac{e_1y_6}{Y_3}) \\ (Y_1,-\frac{(e_5+e_1)y_5}{Y_1}) & (X,0) & (Y_4,\frac{e_3y_7}{Y_4}) & (Y_5,\frac{(e_8+e_3)y_8}{Y_5}) \\ (Y_2,0) & (Y_4,-\frac{e_3y_7}{Y_4}) & (X,0) & (Y_6,0) \\ (Y_3,-\frac{e_1y_6}{Y_3}) & (Y_5,-\frac{(e_8+e_3)y_8}{Y_5}) & (Y_6,0) & (X,0) \end{bmatrix}$$

where

$$Y_1 = y_5 + \frac{y_1y_2}{y_1+y_2+y_3+y_4} \qquad Y_2 = \frac{y_1y_3}{y_1+y_2+y_3+y_4}$$

$$Y_3 = y_6 + \frac{y_1y_4}{y_1+y_2+y_3+y_4} \qquad Y_4 = y_7 + \frac{y_3y_4}{y_1+y_2+y_3+y_4}$$

$$Y_5 = y_8 + \frac{y_2y_4}{y_1+y_2+y_3+y_4} \qquad Y_6 = y_{10} + \frac{y_3y_4}{y_1+y_2+y_3+y_4}$$

(c)

Example 1

IV. REDUCTION OF ACTIVE NETWORKS

The preceeding discussion has been restricted to networks in which all voltage sources were independent. On the other hand, suppose that several sources of a dependent or controlled nature are present in a given network. That is, these sources are functions of voltages and currents throughout the network. Before a reduction process can be applied to such a network, it is necessary to determine what effect the elimination of a vertex will have on a source dependent on a current flowing into that vertex.

Certain restrictions can be imposed on the active elements. Once again, all sources will be represented by their voltage equivalent form. Controlled voltage sources will be assumed to be linear functions of the currents in the network. Such currents will be termed control currents. Therefore, any voltage source may be written

$$E_{ij} = \sum_{k,m} Z_{km} I_{km} + V_{ij} \tag{9}$$

where I_{km} is the current in the branch between vertices v_k and v_m (I_{km} is positive for $k < m$), Z_{km} is the linear dependency of E_{ij} on I_{km}, and V_{ij} is an independent voltage source.

The generalized active network will thus consist of an arbitrary number of passive and active branches. It can be shown that any network whose currents have a unique solution can be represented by an equivalent one with no parallel branches [12]. Hence, we will assume that our network is given in that form. The method of analysis is similar to the passive procedure. Voltage sources, of both the dependent and independent varieties, can be pushed through vertices to clear a particular vertex for a star mesh transformation. If no control currents flow into a cleared vertex, it may immediately be reduced. At this point, the presence of control currents elsewhere in the network necessitates a bookkeeping procedure to preserve proper voltage-current relationships.

Purely passive vertices may be reduced until a vertex v_i, with an incident control current I_{ij} is reached. V_i can not be directly reduced without destroying the relationship between this current and some voltage source in the network. However, a redefinition of I_{ij}

eliminates the dependency at vertex v_i. Let the vertices adjacent to v_j be $v_{j1}, v_{j2}, \ldots, v_{jq}$. Furthermore, let j be less than j1, j2,,, jk and the greater that jk+1, jk+2, , , , jq. The currents in the branches incident at v_j are then $I_{j,j1}, I_{j,j2}, , , , I_{j,jk}, I_{jk+1,j}, I_{jk+2,j}, , , I_{jq,j}$. I_{ij} can be written as the sum of the currents leaving v_j.

$$I_{ij} = \sum_{m=1}^{k} I_{j,jm} - \sum_{m=k+1}^{q} I_{jm,j} \tag{10}$$

The right hand side of (10) can now be substituted for I_{ij} wherever I_{ij} appears. For example

$$E_{rs} = Z_{ij}I_{ij} + V_{rs} = Z_{ij} \sum_{m=1}^{k} I_{j,jm} - Z_{ij} \sum_{m=k+1}^{q} I_{jm,j} + V_{rs} \tag{11}$$

Thus, all the dependencies on v_i can be removed in this fashion and the star mesh transformation applied.

Figure 6 illustrates the above situation for a typical vertex v_j of a network with no parallel branches. It is important to note that before the vertex reduction technique can be applied (for either the passive or active cases), the network must be specified or placed in a non parallel form. Each time a vertex is reduced by a star mesh transformation, the network will lose its non parallel form. Hence, at each stage it must be returned to that form.

After the transformation is applied, parallel branches may be absorbed into the network in the following manner. During the analysis procedure, only passive branches can be generated. At each stage these new branches must be combined to return the network to its non-parallel form. A passive branch may be generated in parallel with any one of a variety of element configurations. The most general case is where the configuration contains both a passive branch and a voltage source dependent on both the current through it and others throughout the network. Figure 7 illustrates this situation.

Let $Z_2I_2+Z_3I_3+++Z_qI_q + V_{11} = V_a$. The dependency of V_1 on I_1 can be removed by letting $Z_a' = Z_a + Z_1$. If $Z_a' = 0$, the problem reduces to a voltage source in parallel with a passive element. For $Z_a' \neq 0$, a simple set of source transformations produces a non-parallel network as in Figure 8. The current I_1 in the original branch may control a voltage elsewhere in the network. In this case, I_1 must be determined as a function of I_1', the current in the non-parallel circuit. This function then replaces I_1 wherever I_1 appears. A general equation for I_1 is

$$I_1 = \frac{I_1'}{1+Y_b/Y_a+Z_1Y_b} - \frac{Y_bV_a}{1+Y_b/Y_a+Z_1Y_b} \qquad (12)$$

where

$$V_a = Z_2I_2 + \ldots + Z_qI_q + V_{11} \qquad (13)$$

is the sum of the independent voltage component of V_1 and the dependency on all of the currents elsewhere in the network.

Similar to the passive technique, a systematic vertex reduction procedure may be developed for active networks. However, such a procedure is rather complicated and the reader is referred to [12] for such a development. Figure 9 is an example of the analysis technique for active networks.

V. PROGRAMMING CONSIDERATIONS AND RELATED PROBLEMS

The previous sections have demonstrated a vertex elimination method of analysis which is independent of the sequence in which the vertices are reduced. Until this point, the vertex order has been of no concern. However, from the standpoint of computational efficiency, it could be an extremely important factor.

When a star mesh transformation is performed on a vertex with n incident branches, $n(n-1)/2 - n = n(n-3)/2$ branches are added to the network. Parallel branches must be combined, an operation requiring computer time. In addition, the total number of branches in the network could grow appreciably within the first few vertex reductions. Since the time required to analyze the network can be expected to be proportional to the number of branches in the network, it is desirable to choose a vertex sequence which minimizes the number of branches added. The problem is further complicated by the presence of voltage sources throughout the graph. After several pliars entries, the number of voltage sources in the network could become to equal the number of branches. The identical problem arises when control currents are considered. Thus, an optimum sequence is one which adds the smallest number of voltage sources and control currents to the network. In addition, the best sequence should be computed as a function of computer time and storage requirements. In general, it was not possible to obtain a solution to this problem.

One additional problem which further complicates strategy is that when impedances are multipled together, common factors in the numerator and denominator polynomials of the parameters will not be cancelled. Because of the repeated addition and multiplication of ratios of polynomials, these common factors will be introduced at most stages of the reduction process. The common factors will in general be a function of the values of the network elements. The introduction of common factors is a significant problem because the order of the solution is directly related to computer time, accuracy, and storage requirements. If care is not taken in choosing the proper vertex sequence, the order of the solutions increases very rapidly. In any event, it is necessary to include factoring subroutines in computer programs written to analyze networks. The common factors are then removed as they are formed.

VI. CONCLUSION

This paper has presented a method for the reduction of networks to their equivalent three terminal representation. The method presented has the advantage that it is relatively easy to program and does not require a large number of logical operations. In addition, if the common roots are removed accurately, as they are formed, efficient programming could provide a fast solution while at the same time conserving computer space.

The reduction method is more efficient than a loop equation solution, and requires less storage space than a topological method where the trees are generated individually. In the multiport case, a topological method of analysis requires that the problem be repeated as many times as there are voltage drivers in the network, and then superposition applied. With the reduction method, the network is analyzed only once.

On the negative side, the addition of common factors requires that factoring subroutines be incorporated in the computer program. As with any numerical technique, error will be introduced each time factoring occurs. If the factoring is repeated a large number of times, this error could become significant. In addition, extra time will be consumed in the factoring portion of the program.

What is needed to increase the significance of the results of this paper is some method to either prevent the formation of the common factors, or to remove the common factors without factoring the entire numerator and denominator polynomials. The strategy problem must also be considered in more detail.

ACKNOWLEDGEMENT

The author would like to express his appreciation to Dr. S. L. Hakimi of Northwestern University, whose many useful suggestions contributed to the development of the results of this study.

BIBLIOGRAPHY

1. Seshu, S. and Reed, M. B., Linear Graphs and Electric Networks, Addison-Wesley Publishing Company, Inc., Massachusetts; 1961.
2. MacWilliams, F.S., "Topological Network Analysis as a Computer Program," IRE Trans., Vol. CT-5, pp. 228-29; September, 1958.
3. Hobbs, E.W. and MacWilliams, F.J., "Topological Network Analysis as a Computer Program (discussion)," IRE Trans., Vol. CT-6, pp. 135-36; March, 1959.
4. Mayeda, W. and Van Valkenberg, M.E., "Network Analysis and Synthesis by Digital Computers," IRE Wescon Convention Record, pp. 137-142; 1957.
5. Hakimi, S. L., "On the Trees of a Graph and Their Generation," Journal of the Franklin Institute, Vol. 272, pp. 347-359; November, 1961.
6. Mayeda, W., "Reducing Computing Time in the Analysis of Networks by Digital Computer," IRE Trans., Vol. CT-6, pp. 136-37; March, 1959.

7. Hakimi, S. L. and Green, D. G., "Generation and Realization of Trees and k-Trees," IRE Trans. on Circuit Theory, June, 1964.

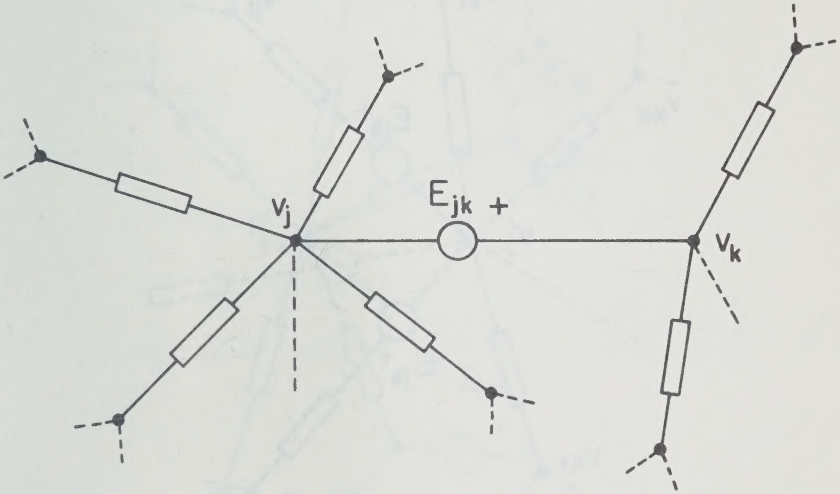
8. Mayeda, W. and Van Valkenberg, M. E., "Analysis of Non Reciprocal Networks by Digital Computer," IRE Convention Record; 1958.

9. Mayeda, W., "Topological Formulas for Non Reciprocal Networks and Networks with Transformers," Proceedings of the NEC, Vol. XIV, pp. 1-13; 1958.

10. Guillemin, E. A., Introductory Circuit Theory, John Wiley and Sons, New York; 1953.

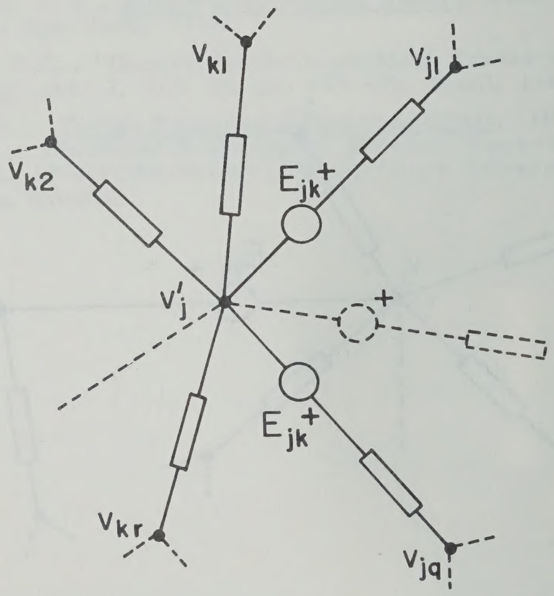
11. Shen, D. W. C., "Generalized Star and Mesh Transformations," Phil. Mag. ser. 7, Vol. 38, pp. 267-275; April, 1947.

12. Frank, H., Vertex Reduction Network Analysis, (Report No. 9, Dept. of Electrical Engineering, Network Theory Group, Grant No. AF-AFOSR-98-63;) Northwestern University, Evanston, Illinois.



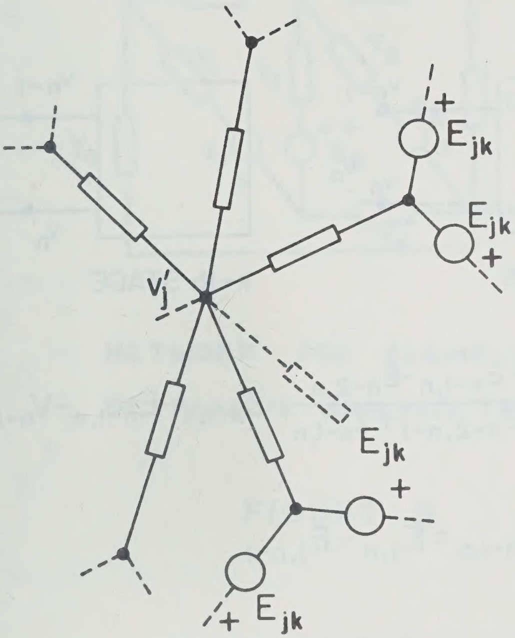
A VERTEX WHICH CANNOT BE DIRECTLY REDUCED BY A STAR-MESH TRANSFORMATION

FIGURE 1


V_{k1}
V_{j1}
V_{k2}
E_{jk} +
V'_j
+
E_{jk} +
V_{kr}
V_{jq}

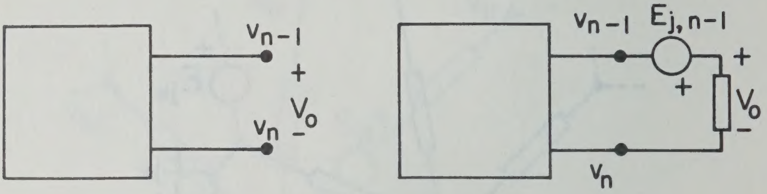
STEP 1 IN THE REDUCTION PROCEDURE

FIGURE 2


V_j'
E_{jk}
+
E_{jk}
+
E_{jk}
+
E_{jk}
+

STEP 2 IN THE REDUCTION PROCEDURE

FIGURE 3


v_{n-1}
+
V_o
v_n
−
v_{n-1}
$E_{j,n-1}$
+
+
V_o
−
v_n

1-st STAGE

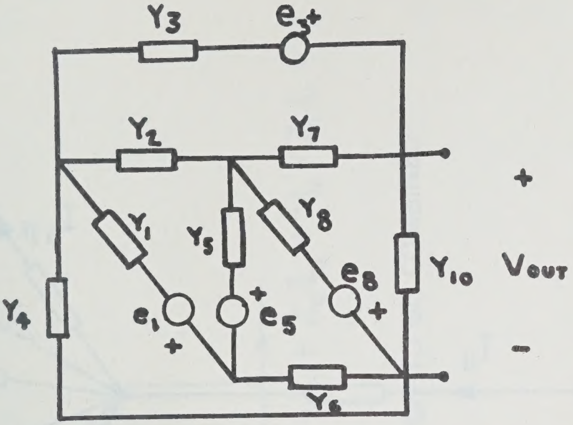
k-th STAGE

$$V_o = \frac{E_{n-2,n-1} + E_{n-1,n} - E_{n-2,n}}{Z_{n-2,n} + Z_{n-2,n-1} + Z_{n-1,n}} Z_{n-1,n} - E_{n-1,n} - V_{n-1,n}$$

$$V_{n-1,n} = E_{j,n} - E_{j,n-1}$$

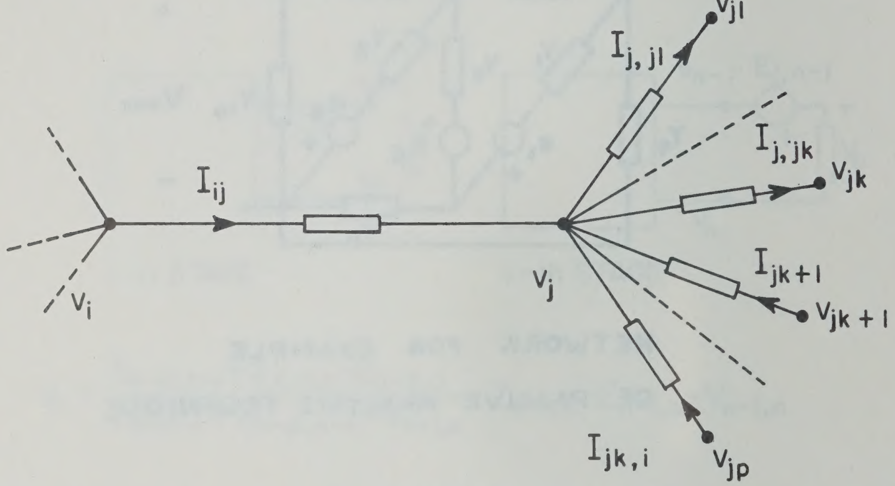
LOADING EFFECT

FIGURE 4


Y3
e3
+
Y2
Y7
Y1
Y5
Y8
Y4
e1
+
+
e5
e8
+
Y10
Vout
+
−
Y6

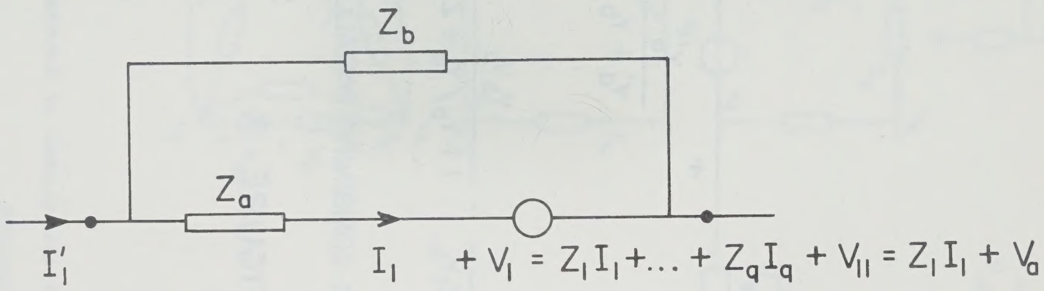
NETWORK FOR EXAMPLE
OF PASSIVE ANALYSIS TECHNIQUE

FIGURE 5


I_{ij}
V_i
V_j
V_{j1}
$I_{j,\,j1}$
$I_{j,\,jk}$
V_{jk}
I_{jk+1}
V_{jk+1}
$I_{jk,\,i}$
V_{jp}

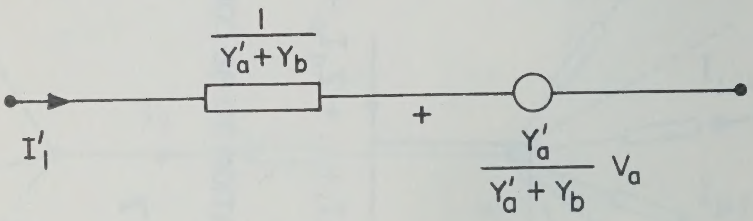
A VERTEX WITH AN INCIDENT CONTROL CURRENT

FIGURE 6


Z_b
Z_a
I_1'
I_1
$+ V_1 = Z_1 I_1 + \ldots + Z_q I_q + V_{11} = Z_1 I_1 + V_a$

GENERAL PARALLEL CONFIGURATION TO BE COMBINED

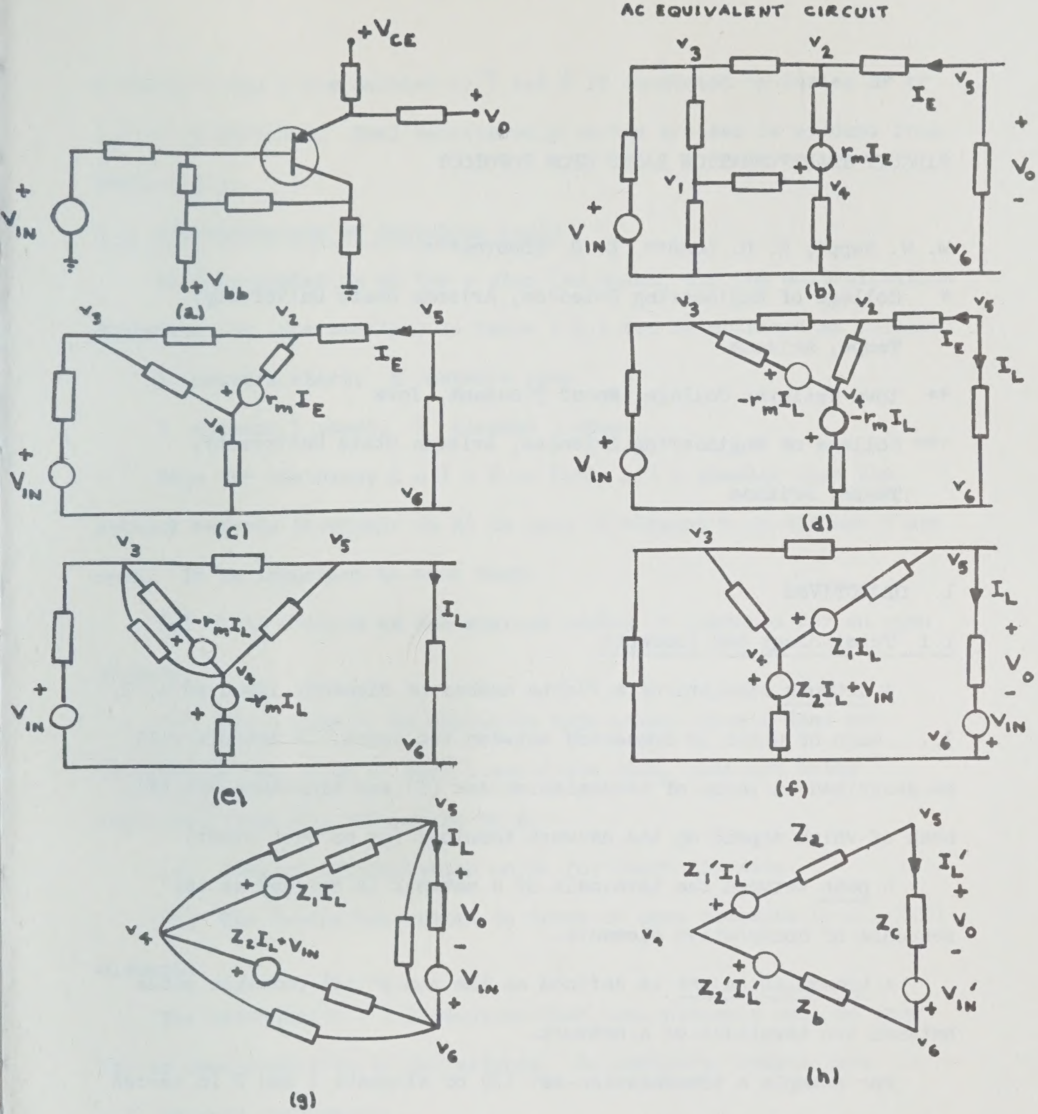
FIGURE 7


$\frac{1}{Y_a' + Y_b}$
I_1'
+
$\frac{Y_a'}{Y_a' + Y_b} V_a$

$$I_1 = \frac{I_1'}{1 + Y_b/Y_a + Z_1 Y_b} - \frac{Y_b V_a}{1 + Y_b/Y_a + Z_1 Y_b}$$

CONFIGURATION AFTER COMBINING PARALLEL BRANCH

FIGURE 8


AC EQUIVALENT CIRCUIT
+VCE
+
Vo
+
VIN
+Vbb
(a)
v3
v2
v5
IE
v1
rmIE
+
v4
+
Vo
-
+
VIN
v6
(b)
v3
v2
v5
IE
+
rmIE
v4
+
VIN
v6
(c)
v3
v2
v5
IE
IL
+
-rmIL
v4
-rmIL
+
+
VIN
v6
(d)
v3
v5
IL
-rmIL
+
v4
-rmIL
+
+
VIN
v6
(e)
v3
v5
IL
+
Vo
-
+
Z1IL
v4
+
Z2IL+VIN
VIN
+
v6
(f)
v5
IL
+
Z1IL
v4
Vo
-
Z2IL+VIN
+
VIN
+
v6
(g)
v5
Za
IL'
Z1'IL'
+
+
Zc
Vo
-
v4
+
Z2'IL'
Zb
VIN'
+
v6
(h)

AN EXAMPLE TO OUTLINE THE ACTIVE ANALYSIS TECHNIQUE

FIGURE 9

CIRCUIT TRANSFORMATION BASED UPON TOPOLOGY

W. W. Happ*, R. D. Lamb**, E. R. Robbins***

* College of Engineering Sciences, Arizona State University,
Tempe, Arizona

** Iowa Wesleyan College, Mount Pleasant, Iowa

*** College of Engineering Sciences, Arizona State University,
Tempe, Arizona

1. OBJECTIVES

1.1 Terminology and Concepts

A network consists of a finite number of elements labelled 1, 2, 3,... each of which is connected between two nodes. A network will be described in terms of transmission-set (I) and hindrance-set (E), both of which depend on the network topology (or on each other).

A path between the terminals of a network is defined as the sequence of consecutive elements.

A transmission-set is defined as the sum of all possible paths between two terminals of a network.

For example a transmission-set (I) of elements 1 and 2 in series is denoted by 12; on the other hand, if 1 and 2 are in parallel, the resulting I-set is denoted by 1 + 2. The multiplication stands for intersection ($\cap$) while the addition sign signifies union (U).

A hindrance-set (E) is defined as the sum of all possible ways of separating two terminals of a network from each other. E-sets must pass through exactly one element of each I-set. The E-set of elements 1 and 2 are denoted by $\bar{1}$ and $\bar{2}$ if connected in series or by $\bar{1}\ \bar{2}$ if in parallel. Dual relationship to the tie set is evident from Table 1.1.1.

1.2 Interpretation of Switching Logic

An interpretation of the + sign (or union) and the multiplication operation (or intersection) in Table 1.1.1 can be obtained as follows:

I network short; E network open

1 element 1 short; $\bar{1}$ element 1 open

Thus the statement $E = \bar{1} + \bar{2}$ in Table 1.1.1 denotes that the network between terminals (A B) is open if element 1 or element 2 are open. It is important to note that:

(a) E is defined as the minimum number of elements for an open circuit.

(b) The + sign is an exclusive type union, thus E does not include the condition if both 1 and 2 are open, this not being consistent with the definition of E.

(c) Similar consideration holds for short circuits.

(d) The E-sets are stated in terms of open elements (e.g., $\bar{1},\bar{2}$) elements.

The statement $E = \bar{1}\ \bar{2}$ requires that both elements must be open for an open condition of the network. No ambiguity exists here.

1.3 Network Constraints

Given a network such as network (d) in Table 1.3.1, the E-set can be expressed as: $E(d) = \bar{5}\ E\ (d,\ \bar{5}) + 5\ E\ (d,\ 5)$

where $E(d,\bar{5})$ is E(d) with element 5 open and E(d,5) is E(d) with element 5 shorted. Note that the E-sets are thus a function of both "open" and "short" elements. Applying this technique to non-planar networks (e) in Table 1.3.1, it is possible to obtain:

$$E(e) = \bar{7}\ 8\ E(e,\bar{7},\bar{8}) + 7\ \bar{8}\ E(e,7,\bar{8}) + \bar{7}\ 8\ E(e,\bar{7},8) + 7\ 8\ E(e,7,8)$$

where $E(e,\bar{7},\bar{8})$ is E(e) with elements 7 and 8 omitted. The $E(e,\bar{7},\bar{8})$ set is referred to as a sub-hindrance-set.

For n constraints m^n sub-sets must be used (each a sub-set of the initial network) where m = 2 if the network has two states, such as open and short. For ternary logic m = 3.

The E-set and I-set of networks of arbitrary complexity can be written down by inspection.

2. MODULAR CODING

Modular coding is a technique of describing properties of a system with a finite number (K) of elements, each element having a finite number (m) of states, where m is referred to as the modulus.

2.1 Terminology and Definitions

The system described on the basis of hindrance-sets (E) and transmission-sets (I) is as follows:

(a) The modulus m = 2 corresponds to two states E_j or I_j for each element j. These are referred to as E-sets and I-sets respectively.

(b) The system is described by two sets $E_s = f(E_j,I_j)$ and $I_s = f(E_j,I_j)$, each being the sum of (E_j,I_j) subsets which are the product of K sets of elements, where K is the number of elements (or literals) in the system.

(c) If the number of "E" states in E_s is K_{EE} and the number of "I" states in E_s is K_{EI}, then $K_{EE} + K_{EI} = K$ for $E_s = f(E_j,I_j)$ and, similarly, $K_{IE} + K_{II} = K$ for $I_s = f(E_j,I_j)$, that is, both E_s and I_s have K literals in each term. Refer to table 2.1.1.

(d) For a single element $K_{EE} = K_{II} = K = 1$. If an element is added by adding a node, then K_{II} and K_{EI} increase by one.

For a system with internal nodes (N), $N = K_{EI} = K_{II} - 1$.

For a planar system with loops (L), $L = K_{IE} = K = K_{II}$.

(e) For any system L can be obtained from $N + L = K - 1$. Thus, if any two of N, L, or K are known the K_{EE}, K_{II}, K_{IE}, K_{EI} can be determined. These numbers depend only on the structure or topology of the system, not on the value of the elements.

2.2 Network Properties

The linearity of E_s in E_j and I_j requires that there is one term of each element in E_s and in I_s, either E_j or I_j for the element j. Thus for K elements, both E_s and I_s consist of the sum of products each with K terms, and this sum is referred to as a homogeneous function of the sets of E_j and I_j. The requirement of homogeneity may be assumed as an axiom. Alternately, an inductive proof of homogeneity can be obtained from the successive application of the following basic rules:

(a) If two series elements constitute a system, then

$$E_s = E_1 I_2 + I_1 E_2 \qquad I_s = I_1 I_2$$

(b) If two parallel elements constitute a system, then

$$E_s = E_1 E_2 \qquad I_s = I_1 E_2 + E_1 I_2$$

(c) If element j is replaced by a short set $I_j = 1$ and $E_j = 0$. If element j is replaced by an open set $E_j = 1$ and $I_j = 0$.

(d) If $E_s(j')$ denotes E_s with element j short and $E_s(j°)$ denotes E_s with element j open, then $E_s = E_j E_s(j')$ and $I_j E_s(j°)$. Similarly, $I_s = E_j I_s(j') + I_j I_s(j°)$.

Each of these operations (a-d) conserves homogeneity. Since any network can be constructed by successive application of these operations, all states E_s and I_s are homogeneous functions of the K elements.

2.3 Systems with Many Elements

From the basic rules, it follows that for both E_s and I_s the K subscripts of the component elements contain each subscript once and hence two alternative shorthand conventions may be adopted. Either subscripts may be omitted, if the symbols are ordered or subscripts out-of-order will denote cyclic summation.

For example

(a) $I_1E_2I_2$will denote $I_1E_2I_3 + I_1I_2E_3$

(b) $I_1E_1I_3E_3$ will denote $(I_1E_2 + E_1I_2)\ (I_3E_4 + E_3I_4)$

(c) $I_1I_1I_1E_1$ denotes $I_1I_2I_3E_4 + I_1I_2E_3I_4 + I_1E_2I_3I_4 + E_1I_2I_3I_4$

For systems with many numbers of elements, this shorthand forms a useful convention as illustrated by the network examples in Table 1.3.1.

2.4 Conversion Techniques

Conversion from modular coding is accomplished as follows:

(a) To switching logic: For E_s set all $I_j = 1$. For I_s set all $E_j = 1$. After reverting from shorthand to explicit function.

(b) To network function: Set $E_j = Z_j$ and $I_j = 1$, or $I_j = Y_j$ and $E_j = 1$ depending on whether the j element is desired as Z_j or Y_j.

(c) From cut-sets to tie-sets: Apply the De Morgan (M) transformation operator and subsequently the homogeneity requirement (H) operator; for example: If $E_s = E_1E_2$, $I_s = H\ M\ E_s = H(I_1 + I_2) = I_1E_2 + E_1I_2$ or $E_s = M\ H^{-1}I_s = M(I_1 + I_2) = E_1E_2$, then the transforms are $I_s = H\ M\ H^{-1}\ E_s$ and $E_s = H\ M\ H^{-1}I_s$ where $H^{-1}E_s$ implies that all $E_j = 1$.

2.5 Transformation Theory

It is desirable to obtain a procedure for the description of a network (1) by switching logic from an equivalent description, (2) the lumped parameter network function with the same topology. A systematic approach must be of scope similar to transformations converting one domain into another as shown in Table 2.5.1.

3. ILLUSTRATIVE EXAMPLES - ONE PORT

The circuit transformation is used in the determination of input impedance where the first example shows the various parts of the problem solution. A second example for a planar network will be given in tabular form.

3.1 Definitions Used in Illustrated Examples

E	the hindrance set
I	the transmission set
E^*	expanded homogeneous E set
I^*	expanded homogeneous I set
$(E_t^*)j$	the jth 2-tree of E* set
$(E_c^*)j$	the jth 2-cotree of E* set
$(I_t^*)_K$	the Kth tree of I* set
$(I_c^*)K$	the Kth cotree of I* set
(M)	De Morgan's theorem operator
(H)	Homogeneous operator
(=)	Network transformation symbol
E', I'	constraint sets
n_v	number of network vertex
n_b	number of network branches
$(x_1 x_2 \ldots x_{n-2} \bar{x}_{n-1} \bar{x}_n)$	all combinations of 2 nots out of n literals

3.2 Single and Multiple Constraints

To determine the input impedance for a one-port, the network is transformed (=) into a switching network with identical topology, that is, branches and vertices must be related in the same pattern.

For a one-port network, a complete transmission set consists of the sum of all possible paths between the input and output terminals. In practice it is important to perform this operation systematically to insure that no path is omitted. Particularly for a large (i.e., many elements) network the possibility of computational error becomes a limiting factor in enumerating all paths. To overcome this limitation the constraint technique was developed and is recommended for large networks. To illustrate, for a non-planar network there are cross-over branches that make it difficult to obtain the transmission set of the network by the elementary rules outlined above. It is then necessary:

(1) To select the cross-over elements as the constraint branches, or any other branches that tend to complicate the network.

(2) To write the transmission set for the sub-network that results when the constraint branch is open.

(3) To write a similar set when the branch is represented by a short.

Multiple simultaneous constraints can be considered. If n branches are constrained then 2^n sub-networks must be evaluated. The following application will show the use of constraints. Refer to Table 3.2.1.

The complete hindrance set can be written as

$$E = E_{\bar{7}\bar{8}} + E_{7\bar{8}} + E_{\bar{7}8} + E_{78}.$$

After either the E-set or I-set has been found the other set can be obtained by using De Morgan's theorem. Complement each literal, and interchange union (+) and intersection (·) symbols. A computer program is being written to perform De Morgan's theorem on large network problems.

3.3 Expansion to Standard Sums

The I-set can now be multiplied out so that no factored terms exist. By standard methods the various terms can be expanded into a standard sum. Only certain of the standard sum terms are desired, namely those terms that can be a tree set. The tree set consists of the terms that have the correct number of true literals N_v-1 in number. In the expanding operation the exclusive "or" process is used, so that all identical terms will be removed from the standard sum expression. The set of the desired transmission terms is marked as the I* set. For the example, the I* set requires 5 tree branches and 3 cotree branches. See Table 3.3.1.

The E-set can also be expanded into a standard sum expression where the special E*-set is a list of all the tree sets for the sub-network after the input terminals are cut apart by a cutting surface. For the example there are 4 tree branches (true literals) and 4 cotree branches (not literals) for the two cut sub-networks. See Table 3.3.2.

After taking the required permutations and eliminating some of the terms by the exclusive "or" operation, the tree and cotree sets can be found.

3.4 Selection of Trees and Co-Trees

The I-set is the transmission set between the input terminals. The I*-subset is the expanded sub-set of branches that gives the tree-set (I_t*) of the network (true literals) and the cotree-set (I_c*) of the network (not literals).

The E-set (hindrance-set) has two ways of interpretating the results.

(1) The E-set must cut the terminals apart which leaves two

sub-networks. The E*-set must not only cut the input terminals apart, but then produce trees (E_t*) in the two sub-networks.

(2) The other way of analysis is from the topology of the cut surfaces. The cutting surface not only cuts the input terminals apart, but then makes other cuts until a cutting surface tree exists. The set of these cut trees are equal to the 2-cotree set (E_c*).

3.5 Transforms for Impedance

Once trees and cotrees of the I-set have been found, then the input impedance can be found using Kirchhoff's third and fourth law based upon the network topology.

$$Z_{in} \;(=)\; \sum_{n=1}^{j} (E_c{*})_j \Big/ \sum_{n=1}^{K} (I_c{*})_K \qquad (1)$$

3.6 Planar Network Example

A planar example is given where the constraint method is not needed and where De Morgan's Theorem is used to generate either the E or I-set from the other set. Refer to Table 3.6.1. The driving-point impedance is obtained by using expression 1 from section 3.5.

4. TWO-PORT THEORY

4.1 Constrained Two-Ports

To evaluate the performance of a two-port network, it is instructive to impose a constraint on one port and obtain a measurement on the terminals of the other port. The constrained port may be open-circuited (i=0) or short-circuited (e=0). The measurement on the terminals of the other port may be a transmission I = (I,C) or a hindrance E = (E,C) where C is the constraint. Constrained transmittance and hindrance sets are referred to as deviations. For example, E =(E,C) is the deviation of E with respect to C.

All possible measurements of a two-port are tabulated in Table 4.1.1 with sign convention defined in Section 4.2. For example, a measurement taken across the input terminals (e_1) with the output open circuited ($i_2 = 0$) yields the network parameter $g = (e_1, i_2)$ for the two-port while the procedure for (i_2, e_1) yields the parameter ($-g$) which is the same in magnitude but negative in sign.

Tables 4.1.2 and 4.1.3 summarize the results obtained by applying measurements and constraints on simple T and π networks, respectively. Note that the transmission or hindrance sets yield the same value, except for sign, and therefore can be used interchangeably.

Table 4.1.4 defines deviations in terms of measurements and constraints and constitutes a generalization of the results given in Tables 4.1.2 and 4.1.3.

4.2 Sign Convention

To assign positive or negative signs to deviations (M,C), the sense of voltages and currents of the two-port is defined from Figure 4.2.1 and the following considerations:

(a) The positive sign of the deviation is defined to be the I or E set from the port where the measurement is taken to the constrained port. The direction from the input port to the output port is taken as positive.

(b) Since both the a and b deviations are defined at the same port, the sign convention is determined by the reference direction of current. That is, the I-sets are positive and the E-sets are negative. Table 4.2.2 tabulates the sign convention. The first column is thus defined as positive deviations.

4.3 Deviations(1)

Using the concepts introduced in Section 4.1, it is possible to express network parameters in deviation notation as found in the literature.$^{(2)}$ Table 4.3.1 tabulates these relationships. For example, the network function h_{21} can be expressed as the ratio of the deviations a and h or

$$h_{21} = \frac{a}{h} = \frac{(i_2,e_2)}{(i_1,e_2)} = \frac{i_2}{i_1}\Bigg|_{e_2 = 0}$$

5.0 DISCUSSION$^{(3),(4),(5),(6)}$

Seshu$^{(7)}$ recognized the close relationship between driving-point admittance (impedance) of a two terminal network and the transmission (hindrance) of a switching network. Further he established that the carry-over of concepts and techniques from network theory made it possible to assist in the synthesis of switching circuits.

This investigation compares the logic upon which network theory and switching is based and to distinguish common aspects from diverse aspects. Common features were found in the topological structures, diverse features in the association of logical operation associated with the interaction of components contained in the structure.

Like Seshu's work, the approach here derives procedures to convert functions from one type of circuit to the other. However the method used differs from Seshu by basing derivations upon the distinct type of logic of each scheme, leading to systems of cut-set and tie-sets. The selection of appropriate cut-set and tie-set become the operation to be examined most critically. Thus with the exception of the identification of common topology, the results of this investigation are new results, both in methods and in specific rules and formulae. Further, the results are not limited to circuits realizable by switching elements and can be extended to other important classes of circuits not included in the aforementioned investigation, specifically

(a) magnetically and capacitively coupled circuits$^{(8)}$

(b) circuit with active elements$^{(9)}$

(c) two-ports and more generally multiports$^{(10)}$

(d) bipolar circuits$^{(11)}$

(e) piece-wise linear circuits (12).

Investigations into each type of circuit listed are topics of separate investigations nearing completion.

G. REFERENCES

1) W. W. Happ, "Dynamic Characteristics of Four-Terminal Networks", IRE Convention Record, 1954.

2) T. R. Nisbet and W. W. Happ, "The calculus of deviations applied to transistor circuit and network analysis", J. Brit. I.R.E., 21, pp. 437-50, May 1961.

3) S.L. Hakimi, "On Trees of a Graph and their Generation", Journal of the Franklin Institute, Vol. 272, pp. 347-359, November, 1961.

4) H. M. Trent, "Note on the Enumeration and Listing of all Possible Trees in a Connected Linear Graph," Proc. Natl. Acad. Sci. U.S., pp. 1004-1007, October, 1954.

5) S. L. Hakimi, "Graphs with Two Kinds of Elements", Journal of the Franklin Institute, Vol. 270, pp. 451-467, 1960

6) Wataru Mayeda and M. E. Van Valkenburg, "Network Analysis and Synthesis by Digital Computer", IRE Wescon Convention Record, Part 2, pp. 137-144, 1957

7) Seshu, S., "On Electric Circuits and Switching Circuits", Interim Technical Report No. 2, Contract No. DA-11-022-ORD-1983, University of Illinois, Engineering Station, Urbana, Illinois; June 15, 1956.

8) J. B. Compton and W. W. Happ, "Capacitive Interactions in Film-Type Circuits", Proc. Natl. Elec. Conf., 19, p. 721-729, Oct. 1963.

9) W. W. Happ, "Applications of Flow Graph Techniques to the Solution of Reliability Problems", in Physics of Failure in Electronics, Vol. II, Spartan Books Inc., 1964.

10) W. W. Happ and P. S. Castro, "Distributed Parameter Circuit Design Techniques", Proc. Natl. Electronics Conf., 17, pp. 44-70, 1961.

11) J. B. Compton and W. W. Happ, "Flowgraph Models of Integrated Circuits and Film-type Systems," Proc. International Conf. Aerospace Electrotechnology (IEEE), April 1964.

12) E. R. Robbins, D. E. Moody and W. W. Happ, "Topological Techniques for Modeling Piecewise-Linear Systems," Submitted to Trans. IEEE - PTGANE.

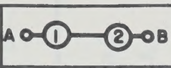
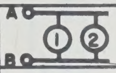
	Topology	Switching Circuit		
(I)	Tie-set	Transmission	12	1 + 2
(E)	Cut-set	Hindrance	$\bar{1} + \bar{2}$	$\bar{1}\,\bar{2}$

TABLE 1.1.1

Modular Coding	Circuit Designation
K_{EE}	2-cotree
K_{EI}	2-tree
K_{II}	tree
K_{IE}	cotree

TABLE 2.1.1

Transform		
Laplace Transform	Time Domain	Frequency Domain
Fourier Transform	Time Variation	Frequency Spectrum
Logarithms	Linear	Exponential
Conformal Mapping	z-Plane	w-Plane
Generating Function	Probability Distribution	Cumulant Moments
Modular Coding	Digital	Lumped Parameter

Table 2.5.1

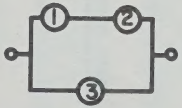
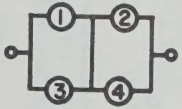
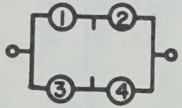
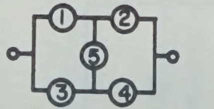
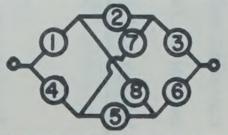
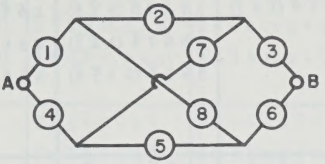
Circuit Examples	Switching Logic		Homogeneous Logic	
	Hindrance Set E-set	Transmission Set I-set	Hindrance Set E-set	Transmission Set I-set
(a) 	$\bar{1}\,\bar{3}+\bar{2}\,\bar{3}$	$1\,2+3$	$E_{(a)}=E_1I_1E_3=$ $E_1I_2E_3+I_1E_2E_3$	$I_{(a)}=I_1I_1E_1=$ $I_1I_2E_3+I_1E_2I_3$ $+E_1I_2I_3$
(b) 	$\bar{1}\,\bar{3}+\bar{2}\,\bar{4}$	$1\,2+1\,4+3\,4+2\,3$	$E_{(b)}=I_1E_1E_1E_1=$ $I_1E_2E_3E_4+E_1I_2E_3E_4$ $+E_1E_2I_3E_4+E_1E_2E_3I_4$	$I_{(b)}=E_1E_2I_1I_2=$ $E_1E_2I_3I_4+I_1E_2E_3I_4$ $+E_1I_2I_3E_4+I_1I_2E_3E_4$
(c) 	$\bar{1}\,\bar{3}+\bar{1}\,\bar{4}+\bar{2}\,\bar{3}+\bar{2}\,\bar{4}$	$1\,2+3\,4$	$E_{(c)}=$ $I_1E_1I_3E_3$	$I_{(c)}=$ $I_1I_1I_1E_1$
(d) 	$\bar{c}\,\bar{5}+\bar{b}\,5$ (b and c refer to the circuit examples)	$c\,\bar{5}+b\,5$	$E_{(c)}E_5+E_{(b)}I_5$	$I_{(c)}E_5+I_{(b)}I_5$
(e) 	$\bar{7}\,\bar{8}\,E\,(e,\bar{7}\,\bar{8})$ $+7\,\bar{8}\,E\,(e,7,\bar{8})$ $+\bar{7}\,8\,E\,(e,\bar{7},8)$ $+7\,8\,E\,(e,7,8)$	$\bar{7}\,\bar{8}\,I\,(i,\bar{7}\,\bar{8})$ $+7\,\bar{8}\,I\,(i,7\,\bar{8})$ $+\bar{7}\,8\,I\,(i,\bar{7}\,8)$ $+7\,8\,I\,(i,7,8)$	This expression would be built up from a dictionary of simple circuit-E and I sets.	

Table 1.3.1

CIRCUIT



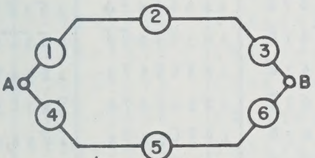
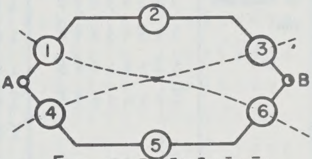
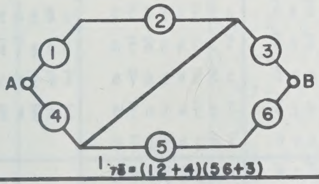
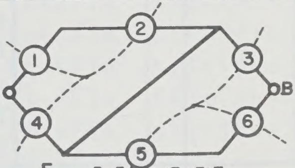
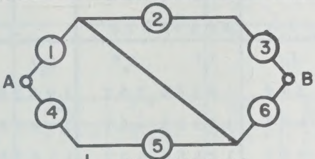
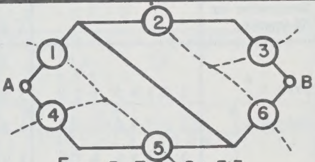
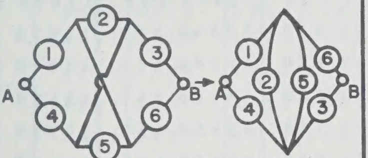
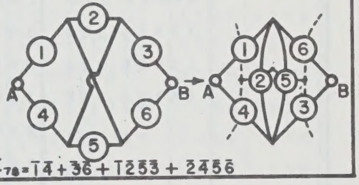
CONSTRAINT	I SET	E SET
$\bar{7}\bar{8}$	 $I_{\bar{7}\bar{8}} = 123 + 456$	 $E_{\bar{7}\bar{8}} = (\bar{1}+\bar{2}+\bar{3})(\bar{4}+\bar{5}+\bar{6})$
$7\bar{8}$	 $I_{7\bar{8}} = (12+4)(56+3)$	 $E_{7\bar{8}} = (\bar{1}+\bar{2})\bar{4}+(\bar{5}+\bar{6})\bar{3}$
$\bar{7}8$	 $I_{\bar{7}8} = (1+45)(23+6)$	 $E_{\bar{7}8} = (\bar{4}+\bar{5})\bar{1}+(\bar{2}+\bar{3})\bar{6}$
78	 $I_{78} = 16+34+13(5+2)+46(5+2)$	 $E_{78} = \bar{1}\bar{4}+\bar{3}\bar{6}+\bar{1}\bar{2}\bar{5}\bar{3}+\bar{2}\bar{4}\bar{5}\bar{6}$

Table 3.2.1

Sub-set

Constraint	$I_{\bar{7}\bar{8}}$	$I_{7\bar{8}}$	$I_{\bar{7}8}$	I_{78}
From Graph **I***	1 2 3 $\bar{7}$ $\bar{8}$ (4 5 $\bar{6}$) 4 5 6 $\bar{7}$ $\bar{8}$ (1 2 $\bar{3}$)	1 2 5 6 7 $\bar{8}$ ($\bar{3}$ $\bar{4}$) 1 2 3 7 $\bar{8}$ (4 $\bar{5}$ $\bar{6}$) 4 5 6 7 $\bar{8}$ (1 $\bar{2}$ $\bar{3}$) 3 4 7 $\bar{8}$ (1 2 $\bar{5}$ $\bar{6}$)	1 2 3 $\bar{7}$ 8 (4 $\bar{5}$ $\bar{6}$) 1 6 $\bar{7}$ 8 (2 3 $\bar{4}$ $\bar{5}$) 2 3 4 5 $\bar{7}$ 8 ($\bar{1}$ $\bar{6}$) 4 5 6 $\bar{7}$ 8 (1 $\bar{2}$ $\bar{3}$)	1 6 7 8 (2 $\bar{3}$ $\bar{4}$ $\bar{5}$) 3 4 7 8 (1 $\bar{2}$ $\bar{5}$ $\bar{6}$) 1 3 5 7 8 ($\bar{2}$ $\bar{4}$ $\bar{6}$) 1 2 3 7 8 ($\bar{4}$ $\bar{5}$ $\bar{6}$) 4 5 6 7 8 ($\bar{1}$ $\bar{2}$ $\bar{3}$) 2 4 6 7 8 ($\bar{1}$ $\bar{3}$ $\bar{5}$)
Expanded in terms of standard sum **I***	1 2 3 4 5 $\bar{6}$ $\bar{7}$ $\bar{8}$ 1 2 3 4 $\bar{5}$ 6 $\bar{7}$ $\bar{8}$ 1 2 3 $\bar{4}$ 5 6 $\bar{7}$ $\bar{8}$ 1 2 $\bar{3}$ 4 5 6 $\bar{7}$ $\bar{8}$ 1 $\bar{2}$ 3 4 5 6 $\bar{7}$ $\bar{8}$ $\bar{1}$ 2 3 4 5 6 $\bar{7}$ $\bar{8}$	1 2 $\bar{3}$ $\bar{4}$ 5 6 7 $\bar{8}$ 1 2 3 4 $\bar{5}$ $\bar{6}$ 7 $\bar{8}$ 1 2 3 $\bar{4}$ 5 $\bar{6}$ 7 $\bar{8}$ 1 2 3 $\bar{4}$ $\bar{5}$ 6 7 $\bar{8}$ 1 $\bar{2}$ $\bar{3}$ 4 5 6 7 $\bar{8}$ $\bar{1}$ 2 $\bar{3}$ 4 5 6 7 $\bar{8}$ $\bar{1}$ $\bar{2}$ 3 4 5 6 7 $\bar{8}$ 1 2 3 4 $\bar{5}$ $\bar{6}$ 7 $\bar{8}$ 1 $\bar{2}$ 3 4 5 $\bar{6}$ 7 $\bar{8}$ $\bar{1}$ 2 3 4 5 $\bar{6}$ 7 $\bar{8}$ 1 $\bar{2}$ 3 4 $\bar{5}$ 6 7 $\bar{8}$ $\bar{1}$ 2 3 4 $\bar{5}$ 6 7 $\bar{8}$ $\bar{1}$ $\bar{2}$ 3 4 5 6 7 $\bar{8}$	1 2 3 4 $\bar{5}$ $\bar{6}$ $\bar{7}$ 8 1 2 3 $\bar{4}$ 5 $\bar{6}$ $\bar{7}$ 8 1 2 3 $\bar{4}$ $\bar{5}$ 6 $\bar{7}$ 8 1 2 3 $\bar{4}$ $\bar{5}$ 6 $\bar{7}$ $\bar{8}$ 1 2 $\bar{3}$ 4 $\bar{5}$ 6 $\bar{7}$ 8 1 $\bar{2}$ 3 4 $\bar{5}$ 6 $\bar{7}$ 8 1 2 $\bar{3}$ $\bar{4}$ 5 6 $\bar{7}$ 8 1 $\bar{2}$ 3 $\bar{4}$ 5 6 $\bar{7}$ 8 1 $\bar{2}$ $\bar{3}$ 4 5 6 $\bar{7}$ 8 $\bar{1}$ 2 3 4 5 $\bar{6}$ $\bar{7}$ 8 1 $\bar{2}$ $\bar{3}$ 4 5 6 $\bar{7}$ 8 $\bar{1}$ 2 $\bar{3}$ 4 5 6 $\bar{7}$ 8 $\bar{1}$ $\bar{2}$ 3 4 5 6 $\bar{7}$ 8	1 2 $\bar{3}$ $\bar{4}$ $\bar{5}$ 6 7 8 1 $\bar{2}$ $\bar{3}$ 4 $\bar{5}$ 6 7 8 1 $\bar{2}$ 3 $\bar{4}$ $\bar{5}$ 6 7 8 1 $\bar{2}$ $\bar{3}$ $\bar{4}$ 5 6 7 8 1 $\bar{2}$ 3 4 $\bar{5}$ $\bar{6}$ 7 8 $\bar{1}$ 2 3 4 $\bar{5}$ $\bar{6}$ 7 8 $\bar{1}$ $\bar{2}$ 3 4 5 $\bar{6}$ 7 8 $\bar{1}$ $\bar{2}$ 3 4 $\bar{5}$ 6 7 8 1 $\bar{2}$ 3 $\bar{4}$ 5 $\bar{6}$ 7 8 1 2 3 $\bar{4}$ $\bar{5}$ $\bar{6}$ 7 8 $\bar{1}$ $\bar{2}$ $\bar{3}$ 4 5 6 7 8 $\bar{1}$ 2 $\bar{3}$ 4 $\bar{5}$ 6 7 8
Exclusive or Discards		1 2 3 4 $\bar{5}$ $\bar{6}$ 7 $\bar{8}$ $\bar{1}$ $\bar{2}$ 3 4 5 6 7 $\bar{8}$	1 2 3 $\bar{4}$ $\bar{5}$ 6 $\bar{7}$ 8 1 $\bar{2}$ $\bar{3}$ 4 5 6 $\bar{7}$ 8	
Tree I_t^* **Cotree I_c^***	I_t^* I_c^* 1 2 3 4 5 6 7 8 1 2 3 4 6 5 7 8 1 2 3 5 6 4 7 8 1 2 4 5 6 3 7 8 1 2 4 5 6 3 7 8 1 3 4 5 6 2 7 8 2 3 4 5 6 1 7 8	I_t^* I_c^* 1 2 5 6 7 3 4 8 1 2 3 5 7 4 6 8 1 2 3 6 7 4 5 8 1 4 5 6 7 2 3 8 1 4 5 6 7 2 3 8 2 4 5 6 7 1 3 8 1 3 4 5 7 2 6 8 2 3 4 5 7 1 6 8 1 3 4 6 7 2 5 8 2 3 4 6 7 1 5 8	I_t^* I_c^* 1 2 3 4 8 5 6 7 1 2 3 5 8 4 6 7 1 2 4 6 8 3 5 7 1 3 4 6 8 2 5 7 1 3 4 6 8 2 5 7 1 2 5 6 8 3 4 7 1 3 5 6 8 2 4 7 2 3 4 5 8 1 6 7 2 4 5 6 8 1 3 7 3 4 5 6 8 1 2 7	I_t^* I_c^* 1 2 6 7 8 3 4 5 1 4 6 7 8 2 3 5 1 3 6 7 8 2 4 5 1 5 6 7 8 2 3 4 1 5 6 7 8 2 3 4 1 3 4 7 8 2 5 6 2 3 4 7 8 1 5 6 3 4 5 7 8 1 2 6 3 4 6 7 8 1 2 5 1 3 5 7 8 2 4 6 1 2 3 7 8 4 5 6 4 5 6 7 8 1 2 3 2 4 6 7 8 1 3 5
Number of Trees per Sub-set	6	9	9	12

Table 3.3.1

Sub-set

Constraint	$E_{\bar{7}\bar{8}}$		$E_{7\bar{8}}$		$E_{\bar{7}8}$		E_{78}	
From Graph E^*	$\bar{1}\bar{4}\bar{7}\bar{8}(2356)+\bar{1}\bar{5}\bar{7}\bar{8}(2346)$ $\bar{1}\bar{6}\bar{7}\bar{8}(2345)+\bar{2}\bar{4}\bar{7}\bar{8}(1356)$ $\bar{2}\bar{5}\bar{7}\bar{8}(1346)+\bar{2}\bar{6}\bar{7}\bar{8}(1345)$ $\bar{3}\bar{4}\bar{7}\bar{8}(1256)+\bar{3}\bar{5}\bar{7}\bar{8}(1246)$ $\bar{3}\bar{6}\bar{7}\bar{8}(1245)$		$\bar{1}\bar{4}7\bar{8}(2 3 5 \bar{6})$ $\bar{2}\bar{4}7\bar{8}(1 3 5 \bar{6})$ $\bar{3}\bar{5}7\bar{8}(1 2 4 \bar{6})$ $\bar{3}\bar{6}7\bar{8}(1 2 4 5)$		$\bar{1}\bar{4}\bar{7}8(2 3 5 \bar{6})$ $\bar{1}\bar{5}\bar{7}8(2 3 4 \bar{6})$ $\bar{2}\bar{6}\bar{7}8(1 3 4 \bar{5})$ $\bar{3}\bar{6}\bar{7}8(1 2 4 \bar{5})$		$\bar{1}\bar{4}78(2 3 \bar{5} \bar{6})$ $\bar{3}\bar{6}78(1 2 \bar{4} \bar{5})$ $\bar{1}\bar{2}\bar{3}\bar{5}78(4 6)$ $\bar{2}\bar{4}\bar{5}\bar{6}78(1 3)$	
Expanded in terms of standard sum E^*	$\bar{1}23\bar{4}56\bar{7}\bar{8}$ $\bar{1}234\bar{5}6\bar{7}\bar{8}$ $\bar{1}2345\bar{6}\bar{7}\bar{8}$ $1\bar{2}3\bar{4}56\bar{7}\bar{8}$ $1\bar{2}34\bar{5}6\bar{7}\bar{8}$ $1\bar{2}345\bar{6}\bar{7}\bar{8}$ $12\bar{3}\bar{4}56\bar{7}\bar{8}$ $12\bar{3}4\bar{5}6\bar{7}\bar{8}$ $12\bar{3}45\bar{6}\bar{7}\bar{8}$		$\bar{1}\bar{2}3\bar{4}567\bar{8}$ $\bar{1}2\bar{3}\bar{4}567\bar{8}$ $\bar{1}23\bar{4}\bar{5}67\bar{8}$ $\bar{1}23\bar{4}5\bar{6}7\bar{8}$ $\bar{1}\bar{2}3\bar{4}567\bar{8}$ $1\bar{2}\bar{3}\bar{4}567\bar{8}$ $1\bar{2}3\bar{4}\bar{5}67\bar{8}$ $1\bar{2}3\bar{4}5\bar{6}7\bar{8}$ $\bar{1}2\bar{3}4\bar{5}67\bar{8}$ $1\bar{2}\bar{3}4\bar{5}67\bar{8}$ $12\bar{3}\bar{4}\bar{5}67\bar{8}$ $12\bar{3}4\bar{5}\bar{6}7\bar{8}$ $\bar{1}2\bar{3}45\bar{6}7\bar{8}$ $1\bar{2}\bar{3}45\bar{6}7\bar{8}$ $12\bar{3}\bar{4}5\bar{6}7\bar{8}$ $12\bar{3}4\bar{5}\bar{6}7\bar{8}$		$\bar{1}\bar{2}3\bar{4}56\bar{7}8$ $\bar{1}2\bar{3}\bar{4}56\bar{7}8$ $\bar{1}23\bar{4}\bar{5}6\bar{7}8$ $\bar{1}23\bar{4}5\bar{6}\bar{7}8$ $\bar{1}\bar{2}34\bar{5}6\bar{7}8$ $\bar{1}2\bar{3}4\bar{5}6\bar{7}8$ $\bar{1}23\bar{4}\bar{5}6\bar{7}8$ $\bar{1}234\bar{5}\bar{6}\bar{7}8$ $\bar{1}\bar{2}345\bar{6}\bar{7}8$ $1\bar{2}\bar{3}45\bar{6}\bar{7}8$ $1\bar{2}3\bar{4}5\bar{6}\bar{7}8$ $1\bar{2}34\bar{5}\bar{6}\bar{7}8$ $\bar{1}2\bar{3}45\bar{6}\bar{7}8$ $1\bar{2}\bar{3}45\bar{6}\bar{7}8$ $12\bar{3}\bar{4}5\bar{6}\bar{7}8$ $12\bar{3}4\bar{5}\bar{6}\bar{7}8$		$\bar{1}\bar{2}\bar{3}\bar{4}5678$ $\bar{1}\bar{2}3\bar{4}\bar{5}678$ $\bar{1}\bar{2}3\bar{4}5\bar{6}78$ $\bar{1}2\bar{3}\bar{4}\bar{5}678$ $\bar{1}2\bar{3}\bar{4}5\bar{6}78$ $\bar{1}23\bar{4}\bar{5}\bar{6}78$ $\bar{1}\bar{2}\bar{3}45\bar{6}78$ $\bar{1}2\bar{3}\bar{4}5\bar{6}78$ $\bar{1}2\bar{3}4\bar{5}\bar{6}78$ $1\bar{2}\bar{3}\bar{4}5\bar{6}78$ $1\bar{2}\bar{3}4\bar{5}\bar{6}78$ $12\bar{3}\bar{4}\bar{5}\bar{6}78$ $\bar{1}\bar{2}\bar{3}4\bar{5}678$ $1\bar{2}3\bar{4}\bar{5}\bar{6}78$	
Exclusive or Discards			$\bar{1}\bar{2}3\bar{4}567\bar{8}$ $12\bar{3}4\bar{5}\bar{6}7\bar{8}$		$\bar{1}23\bar{4}\bar{5}6\bar{7}8$ $1\bar{2}\bar{3}45\bar{6}\bar{7}8$		$\bar{1}2\bar{3}\bar{4}5\bar{6}78$	
	E_t^*	E_c^*	E_t^*	E_c^*	E_t^*	E_c^*	E_t^*	E_c^*
2-Tree E_t^* 2-Cotree E_c^*	2 3 5 6	1 4 7 8	2 5 6 7	1 3 4 8	3 5 6 8	1 2 4 7	5 6 7 8	1 2 3 4
	2 3 4 6	1 5 7 8	2 3 6 7	1 4 5 8	2 5 6 8	1 3 4 7	3 6 7 8	1 2 4 5
	2 3 4 5	1 6 7 8	2 3 5 7	1 4 6 8	2 3 5 8	1 4 6 7	3 5 7 8	1 2 4 6
	1 3 5 6	2 4 7 8	1 5 6 7	2 3 4 8	3 4 6 8	1 2 5 7	2 6 7 8	1 3 4 5
	1 3 4 6	2 5 7 8	1 3 6 7	2 4 5 8	2 4 6 8	1 3 5 7	2 3 7 8	1 4 5 6
	1 3 4 5	2 6 7 8	1 3 5 7	2 4 6 8	2 3 4 8	1 5 6 7	4 5 7 8	1 2 3 6
	1 2 5 6	3 4 7 8	2 4 6 7	1 3 5 8	3 4 5 8	1 2 6 7	2 4 7 8	1 3 5 6
	1 2 4 6	3 5 7 8	1 4 6 7	2 3 5 8	1 3 5 8	2 4 6 7	1 5 7 8	2 3 4 6
	1 2 4 5	3 6 7 8	1 2 6 7	3 4 5 8	1 3 4 8	2 5 6 7	1 4 7 8	2 3 5 6
			2 4 5 7	1 3 6 8	2 4 5 8	1 3 6 7	1 2 7 8	3 4 5 6
			1 4 5 7	2 3 6 8	1 2 5 8	3 4 6 7	4 6 7 8	1 2 3 5
			1 2 5 7	3 4 6 8	1 2 4 8	3 5 6 7	1 3 7 8	2 4 5 6
Number of 2-Trees per Sub-set	9		12		12		12	

TABLE 3.3.2

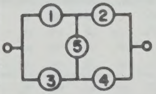
Circuit 	E-set (M) DeMorgan Operator		I-set (M) DeMorgan Operator	
Transmission and Hindrance (sets) E = (M)I = E' I = (M)E = I'	$E=\bar{1}\,\bar{3}+\bar{1}\,\bar{4}\,\bar{5}+\bar{2}\,\bar{3}\,\bar{5}+\bar{2}\,\bar{4}$ $(M)I=(\bar{1}+\bar{2})\,(\bar{1}+\bar{4}+\bar{5})\,(\bar{2}+\bar{3}+\bar{5})\,(\bar{3}+\bar{4})$		$I=1\,2+1\,4\,5+2\,3\,5+3\,4$ $(M)E=(1+3)\,(1+4+5)\,(2+3+5)\,(2+4)$	
Constraint 5 E' = (M)I' = E I' = (M)E' = I	$E'=[\bar{1}\bar{3}+\bar{1}\bar{4}+\bar{2}\bar{3}+\bar{2}\bar{4}]\bar{5}+[\bar{1}\bar{3}+\bar{2}\bar{4}]5$ $(M)I'=[(\bar{1}+\bar{2})(\bar{3}+\bar{4})+5][(\bar{1}+\bar{2})(\bar{1}+\bar{4})(\bar{3}+\bar{4})(\bar{2}+\bar{3})+\bar{5}]$		$I'=[12+34]\bar{5}+[12+14+34+23]5$ $(M)E'=[(1+3)(1+4)(2+3)(2+4)+5][(1+3)(2+4)+\bar{5}]$	
Using E and I terms-obtain E*, I*	$E^*=\bar{1}\,\bar{3}\,(2\,4\,\bar{5})+\bar{1}\,\bar{4}\,\bar{5}\,(2\,3)$ $+\bar{2}\,\bar{3}\,\bar{5}\,(1\,4)+\bar{2}\,\bar{4}\,(1\,3\,5)$		$I^*=1\,2\,(3\,\bar{4}\,\bar{5})+3\,4\,(\bar{1}\,\bar{2}\,\bar{5})$ $+1\,4\,5\,(\bar{2}\,\bar{3})+2\,3\,5\,(\bar{1}\,\bar{4})$	
Expand to Standard Sum E*, I*	$\bar{1}\,2\,\bar{3}\,4\,\bar{5}$ $\;1\,\bar{2}\,\bar{3}\,4\,\bar{5}$ $\;\bar{1}\,2\,3\,\bar{4}\,5$ $\;1\,\bar{2}\,3\,\bar{4}\,\bar{5}$ $\bar{1}\,2\,\bar{3}\,\bar{4}\,5$ $\;1\,\bar{2}\,\bar{3}\,\bar{4}\,5$ $\bar{1}\,\bar{2}\,\bar{3}\,4\,5$ $\;\bar{1}\,\bar{2}\,3\,\bar{4}\,5$		$1\,2\,3\,\bar{4}\,\bar{5}$ $\;1\,\bar{2}\,3\,4\,\bar{5}$ $\;1\,\bar{2}\,\bar{3}\,4\,5$ $\;\bar{1}\,2\,3\,\bar{4}\,5$ $1\,2\,\bar{3}\,4\,\bar{5}$ $\;\bar{1}\,2\,3\,4\,\bar{5}$ $1\,2\,\bar{3}\,\bar{4}\,5$ $\;\bar{1}\,\bar{2}\,3\,4\,5$	
Exclusive or discarded terms	none		none	
Tree I_t^* 2-tree E_t^* Cotree I_c^* 2-cotree E_c^*	E_t^* 24 25 45 14 23 13 15 35	E_c^* 135 134 123 235 145 245 234 124	I_t^* 123 124 125 234 345 145 235	I_c^* 45 35 34 15 12 23 14

Table 3.6.1

Deviation (m,c)	Constraint (C) i_1	i_2	e_1	e_2
Measurement (m)				
I_1	0	z	b	h
I_2	-z	0	-g	a
E_1	-b	g	0	y
E_2	-h	-a	-y	0

Table 4.1.1

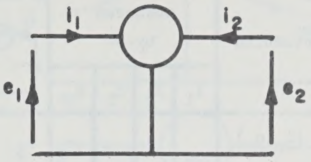


FIGURE 4.2.1

Constrain (= 0) Measure	i_2 E_2	i_1 E_1	e_1 I_2	i_1 E_2	e_1 E_2	i_1 I_2
Constrain (= 0) Measure	e_2 I_2	e_1 I_1	i_2 E_1	e_2 I_1	e_2 E_1	i_2 I_1
Deviation Symbol	(i_2,e_2) a	(i_1,e_1) b	(e_1,i_2) g	(i_1,e_2) h	(e_1,e_2) y	(i_1,i_2) z

Table 4.1.4

Network parameters	Subscript 11 or i	12 or r	21 or f	22 or o	Determinant
a	$-\frac{(g)}{(a)}$	$\frac{(y)}{(a)}$	$-\frac{(z)}{(a)}$	$\frac{(h)}{(a)}$	$\Delta^a = \frac{(b)}{(a)}$
b	$\frac{(h)}{(b)}$	$-\frac{(y)}{(b)}$	$\frac{(z)}{(b)}$	$-\frac{(g)}{(b)}$	$\Delta^b = \frac{(a)}{(b)}$
g	$\frac{(z)}{(g)}$	$-\frac{(b)}{(g)}$	$-\frac{(a)}{(g)}$	$\frac{(y)}{(g)}$	$\Delta^g = \frac{(h)}{(g)}$
h	$\frac{(y)}{(h)}$	$-\frac{(b)}{(h)}$	$+\frac{(a)}{(h)}$	$\frac{(z)}{(h)}$	$\Delta^h = \frac{(g)}{(h)}$
y	$\frac{(h)}{(y)}$	$-\frac{(b)}{(y)}$	$+\frac{(a)}{(y)}$	$\frac{(g)}{(y)}$	$\Delta^y = \frac{(z)}{(y)}$
z	$\frac{(g)}{(z)}$	$\frac{(b)}{(z)}$	$-\frac{(a)}{(z)}$	$\frac{(h)}{(z)}$	$\Delta^z = \frac{(y)}{(z)}$

Table 4.3.1

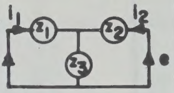
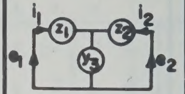
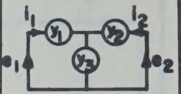
Deviation	Measured by---				Value of Measurement		
	I_1	I_2	E_1	E_2	 i_1 i_2 z_1 z_2 z_3 e	 i_1 i_2 e_1 e_2 z_1 z_2 y_3	 i_1 i_2 e_1 e_2 y_1 y_2 y_3
a (i_2,e_2)		+		-	z_3	1	y_1y_2
b (i_1,e_1)	+		-		z_3	1	y_1y_2
g (e_1,i_2)		-	+		z_1+z_3	z_1y_3+1	$y_1y_2+y_2y_3$
h (i_1,e_2)	+			-	z_2+z_3	z_2y_3+1	$y_1y_2+y_1y_3$
y (e_1,e_2)			+	-	$z_1z_2+z_1z_3+z_2z_3$	$z_1z_2y_3+z_1+z_2$	$y_1+y_2+y_3$
z (i_1,i_2)	+	-			1	y_3	$y_1y_2y_3$

Table 4.1.2

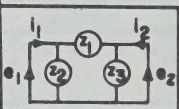
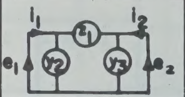
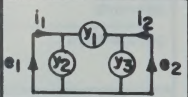
Deviation	Measured by---				Value of Measurement		
	I_1	I_2	E_1	E_2	 i_1 i_2 e_1 e_2 z_1 z_2 z_3	 i_1 i_2 e_1 e_2 z_1 y_2 y_3	 i_1 i_2 e_1 e_2 y_1 y_2 y_3
a (i_2,e_2)		+		-	z_2z_3	1	y_1
b (i_1,e_1)	+		-		z_2z_3	1	y_1
g (e_1,i_2)		-	+		$z_1z_2+z_2z_3$	z_1y_3+1	y_1+y_3
h (i_1,e_2)	+			-	$z_1z_3+z_2z_3$	z_1y_2+1	y_1+y_2
y (e_1,e_2)			+	-	$z_1z_2z_3$	z_1	1
z (i_1,i_2)	+	-			$z_1+z_2+z_3$	$y_2+y_3+z_1y_2y_3$	$y_1y_2+y_1y_3+y_2y_3$

Table 4.1.3

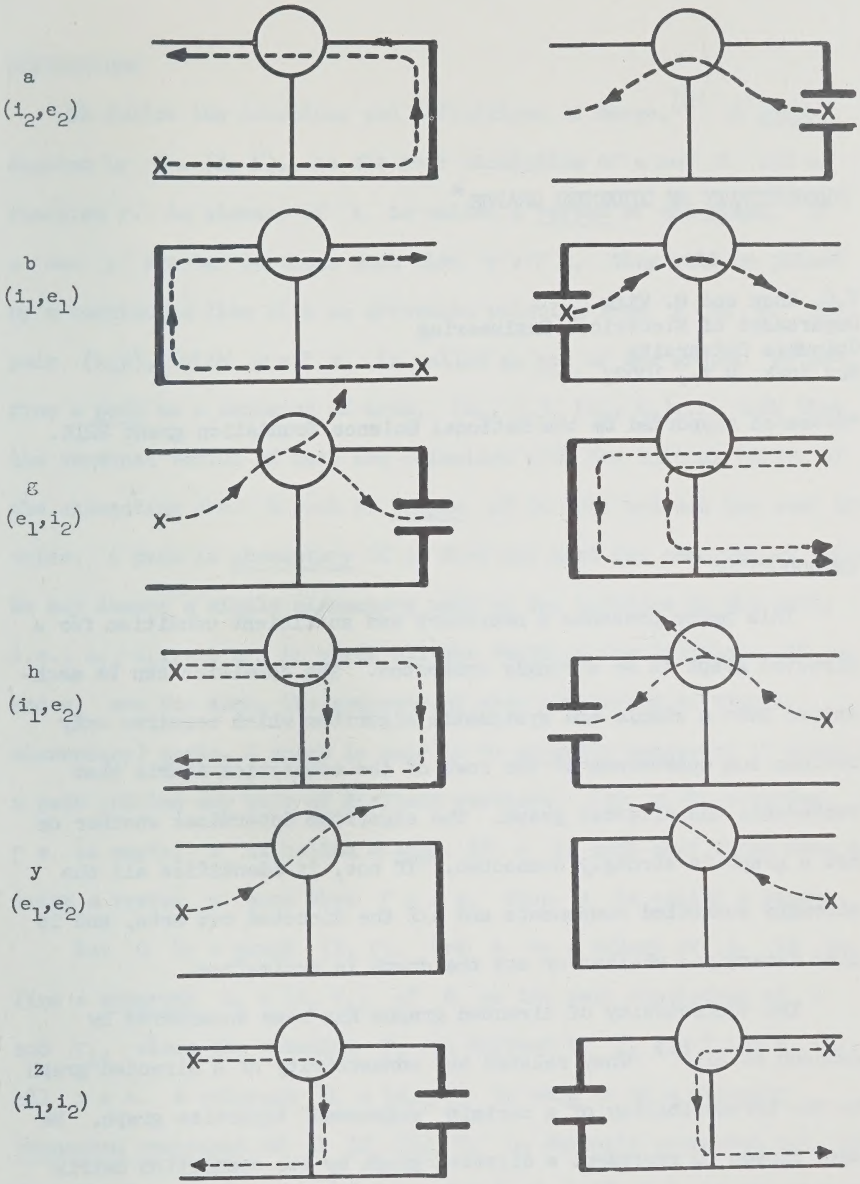

a
(i_2, e_2)
X
X
b
(i_1, e_1)
X
X
g
(e_1, i_2)
X
X
h
(i_1, e_2)
X
X
y
(e_1, e_2)
X
X
z
(i_1, i_2)
X
X

TABLE 4.2.2
Sign Convention for Constrained Two-Ports
In Terms of Transmission and Hindrance Sets

CONNECTIVITY OF DIRECTED GRAPHS*

Y.C. Chen and O. Wing
Department of Electrical Engineering
Columbia University
New York, N.Y., 10027

*Research supported by the National Science Foundation grant 2216.

INTRODUCTION

This paper presents a necessary and sufficient condition for a directed graph to be strongly connected. The condition can be mechanized into a simple and systematic algorithm which requires only Boolean sum operations of the rows of the connection matrix that represents the directed graph. The algorithm determines whether or not a graph is strongly connected. If not, it identifies all the strongly connected components and all the directed cut sets, and it also determines whether or not the graph is cycle-free.

The connectivity of directed graphs has been considered by Johnson et al.(1) They related the connectivity of a directed graph to the irreducibility of a certain "augmented" bipartite graph. We have chosen to represent a directed graph by its connection matrix and to discuss the connectivity of the graph in terms of some easily identifiable properties of the "terminal" connection matrix of the graph, which is readily deducible from the connection matrix. At the end of the paper we demonstrate the equivalence of our results and those obtained by Johnson et al.

DEFINITIONS

We follow the notations and definitions of Berge.(2) A graph, denoted by $G = (X, \Gamma)$, is the pair consisting of a set X and a function Γ. An element of X is called a vertex of the graph. If x and y are two vertices such that $y \in \Gamma x$, they will be joined by a continuous line with an arrowhead pointing from x to y. The pair (x,y), with $y \in \Gamma x$, is called an arc of the graph. We define a path as a sequence of arcs, $(u_1, u_2), (u_2, u_3), \ldots$ such that the terminal vertex of each arc coincides with the initial vertex of the succeeding arc. A path is simple if it does not use the same arc twice. A path is elementary if it does not meet the same vertex twice. We may denote a simple elementary path by the vertices on the path, i.e., $u_1, u_2, \ldots, u_p$ in which all the vertices are distinct. If u_1 and u_p are the same, the sequence of arcs constitute a (simple elementary) cycle. A graph is said to be strongly connected if there is a path joining any pair of distinct vertices. Let x be a vertex. If Γx is empty, x is called a sink. If x is such that there does not exist a vertex y such that $\Gamma y = x$, then x is called a source.

Let G be a graph (X, Γ). Let A be a subset of X. We define a subgraph $G_1 = (A, \Gamma_A)$ of G as the pair consisting of A and Γ_A, where the function Γ_A is defined by $\Gamma_A x = \Gamma x \cap A$, for all $x \in A$. A subgraph $G_1 = (A, \Gamma_A)$ is said to be a strongly connected component of G if, (1) G_1 is strongly connected and, (2) for all B such that $B \neq A$ and $B \supset A$, $G_2 = (B, \Gamma_B)$ is not strongly connected. By definition, an isolated vertex, a sink or a source is considered as a strongly connected component. We define a directed cut set as a set of arcs having the property that its removal from the graph leaves the graph with two disjoint subgraphs such that the terminal vertices of the arcs are in one subgraph and the initial vertices are in the other and such that no proper subset of the set of arcs has this property.

Let a graph G have V vertices. We define the connection matrix, $C = [\ c_{ij}\ ]$, associated with G as a square matrix of order V in which $c_{ij} = 1$ if and only if there exists an arc directed from vertex i to vertex j, and $c_{ij} = 0$ otherwise. We define the terminal connection matrix, $T = [\ t_{ij}\ ]$, associated with a graph G as a square matrix of order V in which $t_{ij} = 1$ if and only if there exists at least one path from vertex i to vertex j and $t_{ij} = 0$ otherwise.(3,4) A vertex of G is called a supervisor of G if it can reach all other vertices of G via some path. A vertex is called a supervisee of G if it can be reached from all other vertices of G via some path.

THEOREM 1.

The element t_{ij} of the terminal connection matrix associated with a graph $G = (X, \Gamma)$ is non-zero if and only if

$$\overset{v-1}{\underset{n=1}{\bar{U}}} \Gamma^n(x_i) \supset x_j \text{ - (1)}$$

where v is the number of vertices of the graph G, Γ is a function mapping X into X, and the functions, $\Gamma^2(x)$, $\Gamma^3(x)$.... are defined by

$$\Gamma^2(x) = \Gamma(\Gamma(x))$$

$$\Gamma^3(x) = \Gamma(\Gamma^2(x)), \text{ etc.}$$

Proof: If $t_{ij} = 1$, by definition there is at least one path from vertex i to vertex j. The set $\overset{v-1}{\underset{n=1}{\bar{U}}} \Gamma^n(x_i)$ contains the set of all vertices that can be reached from x_i via some simple elementary path, the longest of which contains at most v - 1 vertices. Then (1) is true.

If (1) is true, there exists at least one term of the union $\bigcup_{n=1}^{v-1} \Gamma^n (x_i)$ which contains x_j. It means that there exists at least one path connecting vertex x_i to vertex x_j. By definition t_{ij} is a non-zero element.

THEOREM 2.

A vertex x_i is a supervisor of the graph $G = (X, \Gamma)$ if

$$\bigcup_{n=1}^{v-1} \Gamma^n (x_i) \supset x_r \tag{2}$$

where x_r is a supervisor, and $x_i, x_r \in X$.

Proof: By Theorem 1, (2) implies that there exists at least one path from vertex x_i to vertex x_r. Vertex x_r is a supervisor. There exists at least one path from vertex x_r to all other vertices of G. It follows that all other vertices can be reached from x_i, and x_i is a supervisor.

THEOREM 3.

A vertex x_i is a supervisee of the graph $G = (X, \Gamma)$ if

$$\bigcup_{n=1}^{v-1} \Gamma^n (x_e) \supset x_i \tag{3}$$

where x_e is a supervisee, and $x_e, x_i \in X$.

Proof: By Theorem 1, (3) implies that there exists at least one path from vertex x_e to vertex x_i. Vertex x_e is a supervisee. It can be reached from all other vertices of G via some path. It follows that x_i can be reached from all other vertices of G, and x_i is a supervisee.

The next three Theorems (4,5 and 6) follow directly from the definition of the terminal connection matrix.

THEOREM 4.

A graph is strongly connected, if and only if all the elements of the associated terminal connection matrix are non-zero.

THEOREM 5.

A graph is strongly connected if and only if every vertex is simultaneously a supervisor and a supervisee.

Corrollary 1. If a graph contains a vertex which is simultaneously a supervisor and a supervisee, then every vertex is simultaneously a supervisor and a supervisee.

Corollary 2. Let x_r be a supervisor and x_e be a supervisee of a graph G. Where $x_e \neq x_r$. If there is an arc connecting from x_e to x_r, then the graph G is strongly connected.

THEOREM 6.

A graph $G = (X, \Gamma)$ is cycle free if and only if the main diagonal elements of the associated terminal connection matrix are all zeros.

Let G be a graph. Let T be its terminal connection matrix. Let the columns and the rows of T be subject to the same permutation such that identical rows are so arranged that they are adjacent and such that identical columns are so arranged that they are adjacent. Let T' be the terminal connection matrix which has been so arranged. We can state the following theorem.

THEOREM 7.

Let M be a principal minor of maximum order of T'. The vertices associated with the rows (columns) of M belong to a strongly connected component if and only if the elements of M are all non-zero.

Proof: Let A be the set of vertices associated with M. Let

$G_1 = (A, \Gamma_A)$ be a subgraph of G, where $\Gamma_A x = \Gamma x \cap A$, for all $x \in A$.

If all the elements of M are non-zero, by Theorem 4, G_1 is strongly connected. M is a principal minor of maximum order of T'. Then for all B such that $B \neq A$ and $B \supset A$, $G_2 = (B, \Gamma_B)$ is not strongly connected. By definition, G_1 is a strongly connected component of G.

If A is the set of vertices associated with a strongly connected component, by definition $G_1 = (A, \Gamma_A)$ is strongly connected. Let the elements of A be so arranged that their associated rows and columns of the terminal connection matrix are adjacent. The principal minor associated with these rows is in fact the terminal connection matrix of G_1, and by Theorem 4, all elements of the principal minor are non-zero.

ALGORITHM

We now present an algorithm based on these theorems to test the connectivity of a graph. Boolean sum operations of the rows of the connection matrix of the graph are taken to construct the associated terminal connection matrix, (Theorem 1.) From the terminal connection matrix the algorithm determines whether or not a graph is strongly connected. If not, it identifies all the strongly connected components and all the directed cuts sets, and it determines whether or not a graph is cycle-free.

Procedures of Construction of Terminal Connection Matrix.

Start from the first row of the given connection matrix;

a. If all elements of the first row are ones, the vertex associated with this row is a supervisor. Skip to the next row.

b. If the condition of (a) does not exist, but the associated vertex of this row is connected to a supervisor, by Theorem 2, this vertex is a supervisor. Set every element of this row equal to one and work on the next row.

c. If conditions of (a) and (b) do not exist, the following steps are taken:

Step 1. Let $C_{1i}, C_{1j}, \ldots C_{1m}$ be the non-zero elements in the first row of the connection matrix. Add (Boolean sum) the i^{th}, j^{th}, ... and m^{th} rows to the first row. This step corresponds to determine $\Gamma^2 (x_1)$ of (1).

Step 2. Suppose that there are k additional non-zero elements $C_{1p}, C_{1q}, \ldots C_{1r}$ in the first row generated by the first step, then add (Boolean sum) the p^{th}, q^{th}, ... r^{th} rows to the first row. This step corresponds to determine the term $\Gamma^3 (x_1)$ of (1).

Step 3. Repeat step 2 until one or both of the following two conditions occurs:

1. The whole row contains no zero element.

2. No additional non-zero element is generated by the preceeding step.

This procedure is repeated on every row of the connection matrix. The final matrix is the terminal connection matrix associated with the graph G.

If we are only interested in testing whether or not the graph is strongly connected, it is not necessary to construct the associated terminal connection matrix of the graph. A more efficient algorithm can be formulated as follows:

Algorithm

Step 1. Construct the transpose C^T of the connection matrix C.

Step 2. Apply the Boolean sum operation only to the first row of the connection matrix C and the first row of its transpose C^T.

By corollary 1, of Theorem 5, the graph is strongly connected if and only if the first row of C and the first row of C^T after the Boolean sum operation of step 2 contain no zero element.

ILLUSTRATIVE EXAMPLES

Example 1. To test whether or not the graph is strongly connected.

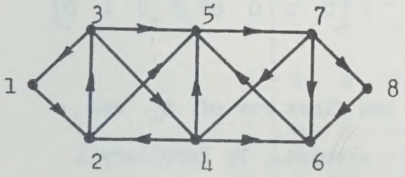


Fig.(1)

$$C_1 = \begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 8 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}$$

Figure (1) is the given graph G_1 and C_1 is its associated connection matrix.

Algorithm

Step 1. Construct C_1^T of C_1.

Step 2. Apply Boolean sum operation to the first row of C and the first row of C_1^T. C_1 and C_1^T becomes respectively C_1' and $(C_1^T)'$ given below. (The Boolean sum operation used in this example is the same as that of construction the terminal connection matrix.)

$$C_1' = \begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 8 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array} \qquad (C_1^T)' = \begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 5 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 6 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 7 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

After the Boolean sum operations the first row of C_1' and the first row of $(C_1^T)'$ contain no zero element. By Corollary 1 of Theorem 5, graph G_1 is strongly connected.

Example 2. To identify all the strongly connected components and all the directed cut sets of the graph.

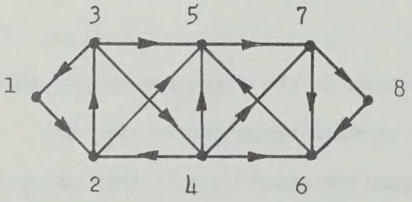


Fig. (2)

$$C_2 = \begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 8 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}$$

Figure (2) is the given graph G_2 and C_2 is its associated connection matrix.

The terminal connection matrix constructed from C_2 is given below:

$$T_2 = \begin{array}{c|cccc|cccc|}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
1 & 1 & 1^* & 1 & 1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1^* & 1 & 1^* & 1 & 1 & 1 \\
3 & 1^* & 1 & 1 & 1^* & 1^* & 1 & 1 & 1 \\
4 & 1 & 1^* & 1 & 1 & 1^* & 1^* & 1^* & 1 \\
\hline
5 & 0 & 0 & 0 & 0 & 1 & 1 & 1^* & 1 \\
6 & 0 & 0 & 0 & 0 & 1^* & 1 & 1 & 1 \\
7 & 0 & 0 & 0 & 0 & 1 & 1^* & 1 & 1^* \\
8 & 0 & 0 & 0 & 0 & 1 & 1^* & 1 & 1 \\
\hline
\end{array}$$

1^*---- Elements of the connection matrix C_2

From T_2, vertices 1, 2, 3, and 4 form one principal minor M_1 of maximum order, while vertices 5, 6, 7, and 8 form another principal minor M_2 of maximum order. M_1 and M_2 contain no zero element. By Theorem 7, vertices 1, 2, 3 and 4 are the vertices of one strongly connected component, and vertices 5, 6, 7 and 8 are the vertices of the other. The arcs (2,5), (3,5), (4,5), (4,6) and (4,7) are the elements of the directed cut set separating the two strongly connected components. All the initial vertices of these arcs are in M_1 and all the terminal vertices of these arcs are in M_2. The removal of these arcs from the graph G_2 leaves the graph with two disjoint subgraphs. Hence these arcs form a directed cut set of the graph G_2.

Example 3. To test whether or not the graph is cycle-free.

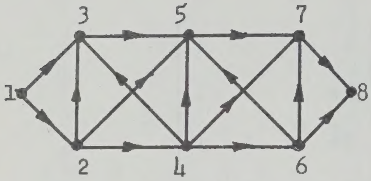


Fig. (3)

$$C_3 = \begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Figure (3) is the given graph G_3 and C_3 is its associated connection matrix.

The terminal connection matrix constructed from the connection matrix C_3 is given below:

$$T_3 = \begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 4 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

The main diagonal elements of T_3 are all zero. By Theorem 6, the graph G_3 is cycle-free.

REMARKS

We remark that Johnson et al$^{(1)}$ obtained a necessary and sufficient condition on a graph G to be strongly connected. The condition is given in terms of the irreducibility of a certain augmented bipartite graph which is derived from G. In terms of our results, the augmented bipartite graph can be represented by an "augmented" connection matrix obtained from the connection matrix of G by letting all the diagonal elements be one. The minimum number of lines that must be drawn either horizontally or vertically, or both, to cover all the non-zero elements of the augmented connection matrix is equivalent to what is called the minmum cover of the augmented bipartite graph. It is clear that if a line is drawn, all zero elements along the line can be replaced by 1's without affecting the number of lines required to cover the matrix. It follows that if the minimum number of lines is equal to the number of vertices of G and if all the lines are either horizontal or vertical, the augmented connection matrix can be replaced by a matrix of 1's and the graph is strongly connected by Theorem 4. Equivalently, it was shown by Johnson et al that if the minimum cover is equal to the number of vertices of G and if one of the two identical sets of vertices of the augmented bipartite graph must be covered, then the graph is strongly connected. We believe our necessary and sufficient condition is more useful in that it can be mechanized into an efficient algorithm applicable to the analysis of graphs having a large number of vertices and arcs.

REFERENCES

1. D.M. Johnson, A.L. Dulmage and N.S. Mendelsohn, "Connectivity and Reducibility of Graphs," Canadian Journal of Mathematics, Vol. 14 (1962), pp. 529-539.

2. C. Berge. Theorie des Graphes et ses Applications. Paris: Durod, 1958.

3. P. Demetriou and O. Wing, "Analysis of Probabilistic Networks," IEEE Transactions on Communication Technology, (September, 1964).

4. P. Demetriou, "Analysis and Synthesis of Probabilistic Networks," Ph.D. Dissertation, Department of Electrical Engineering, Columbia University, July, 1963.

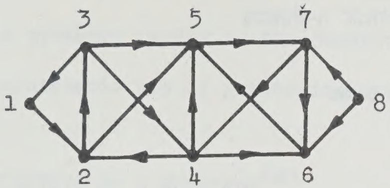


Fig (1)

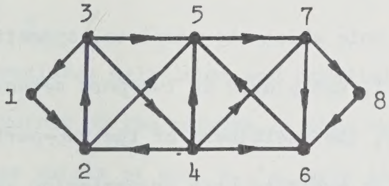


Fig (2)

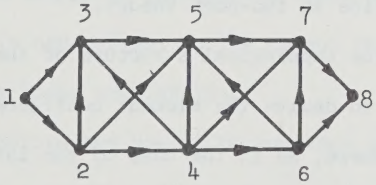


Fig (3)

GENERATION OF SYMMETRIC n-PORTS

B. R. Myers
University of Notre Dame
Notre Dame, Indiana

I. INTRODUCTION

The important role which the complete, symmetric lattice configuration, Figure 1, has played in two-port network theory is well known. Essentially, the usefulness of the two-port lattice derives from the fact that it exhibits both transmission symmetry and port symmetry, as evidenced in the Figure.

The more general and, therefore, more important n-port problem has as yet, however, proved relatively unyielding in the way of canonical or other key n-port structures comparable in significance to that of the lattice in two-port theory.(1)

Suggested by the topological structure of the lattice, one logical problem is to derive the element configuration of a complete, symmetric n-port; where, as in the case of the lattice, the network is assumed to have no internal nodes. For values of n up to perhaps five, the problem can be solved, with a little care, by visual geometry; but beyond that it is impractical to attempt to draw the circuit on paper. Some other, analytic solution is therefore called for.

The problem is solved in this paper via the use of a connection matrix, which is defined in the next section. Ready identification of certain node-pair symmetry cycles in the connection matrix provides a simple and systematic way of generating symmetric n-ports of any order.

II. THE CONNECTION MATRIX OF A NETWORK*(2)

Consider an arbitrary network with v distinct, labelled nodes $v_1, v_2, \dots, v_v$. Let e_{ij} denote the totality of elements connected directly between node v_i and node v_j, where, to avoid duplicity in labelling, it is assumed that

$$1 \leq i < j \leq v$$

This notation also precludes self-loops and parallel elements.

The connection matrix of the network is then the (v-1) by (v-1) upper triangular matrix of the e_{ij}'s, with the rows of the matrix labelled to correspond to the nodes v_i, and the columns to the nodes v_j. For example, the connection matrix of the network which is portrayed by its graph in Figure 2(a) is shown in (b) of the Figure. Clearly, the connection matrix of any network contains all of the topological and other information contained in the actual network. The matrix is non-unique, to the extent that it depends on the order in which the nodes are labelled.

It is both instructive and useful to consider how the node-pair cycle corresponding to a constant distance, k, between nodes appears in the connection matrix; i.e., the node-pair cycle corresponding to

---

*This is actually the same connection matrix as used by N. Nakagawa, "On the evaluation of graph trees and driving point admittance" IRE Trans. on Circuit Theory, CT-5, 2, June, 1958, pp. 122-127

$$(j - i) = k$$

simultaneously with (1)

$$(j-i) = (v - k)$$

where i , j and v are the node label indices, as defined earlier, and the bounds are k are, therefore,

$$1 \leq k \leq (v-1)$$

After the pattern of Riordan$^{(3)}$ there is a cycle of node pairs, given, in terms of the node indices, by

$$(1,1+k;\ 2,2+k;\ \ldots\ ;\ v-k,v;\ v-k+1,1;\ \ldots\ ;\ v,k) \qquad (2)$$

This is a cycle of length v for all k , except when

$$k+1 = v-k+1$$

i.e., when $2k = v$. In which case, v must be even and the cycle length is then $(v-k) = k$.

In the connection matrix of the network, the family of cycles (2), over all k , occupies concentric sets of element positions, each set comprising the two diagonals in the matrix for the two values of k in (1). This is demonstrated in Figure 3 for a 6-node network.

The usefulness of this cycle structure, as it appears in the connection matrix, will become apparent in the next section in determining the connection matrix of a symmetric n-port.

III. GENERATION OF SYMMETRIC n-PORTS

In considering a symmetric n-port with no internal nodes, a convenient geometric representation is to position the totality of $2n$ nodes (the n pairs of terminal nodes) at equidistant intervals around the circumference of a circle. The nodes are then labelled v_1, v_2, ... , v_{2n} in natural order, coincidentally with the n ports around the nodal circle. Thus, node pair (v_1, v_2) constitutes port 1, (v_3,v_4) constitutes port 2, and so on, as depicted in Figure 4 for a 3-port.

With this representation, n-port symmetry requires that the network be symmetrical with respect to each of two sets of n equiangular radii, viz: the n radial axes of port symmetry intersecting the port node-pairs, indicated by solid lines in Figure 4, and the n radial axes of transmission symmetry intersecting adjacent inter-port node pairs, shown by dashed lines in the Figure.

There are, thus, n symmetry diameters in the nodal circle. When n is odd, each diameter comprises a radius of port symmetry which is collinear with a radius of transmission symmetry, as in Figure 4. When n is even, there are n/2 diameters of port symmetry alternating spatially with n/2 diameters of transmission symmetry.

Assuming no self-loops, all elements in the network are connected directly between terminal nodes, there being no internal nodes. Clearly, therefore, the most general network in the class being considered has a non-zero element connected between every distinct pair of nodes. The general n-port is complete in the sense that its graph is a complete graph. There are no zeroes in the connection matrix of a complete network other than the permanent zeroes in the lower triangle below the main diagonal of the matrix. And the total number of elements is equal to the number of combinations of node pairs taken two at a time; which is n(2n-1).

If the n-port is symmetric, then each pair of nodes on the nodal circle is bisected by a symmetry diameter; the same symmetry diameter bisects two diametrically opposite pairs of nodes, in fact. Consequently, the element connections from one node of each adjacent pair to all other nodes on the nodal circle must be symmetric, about the bisecting diameter, with respect to the corresponding connections from the other node of the pair.

Denoting the element connected between nodes v_i and v_j by $e_{i,j}$, this means that

$$e_{i,i+k} = e_{i+1,i+1+2n-k}$$

and (3)

$$e_{i,i+2n-k} = e_{i+1,i+1+k}$$

must hold, simultaneously, for each value of k in the range

$$1 \leq k \leq (2n-1)$$

and for all i from 1 to 2n, inclusive; where, due to cyclic repetition, it is understood in (3) that an index $(2n+p)$, $p>0$, is actually the index p.

A geometric explanation of (3) is apparent from Figures 5(a) and 5(b); Figure 5(a) takes care of cross-over symmetry, and 5(b) of mirror symmetry. Different elements will be located in the positions corresponding to different values of k, in the general instance.

Applying (3) to successive adjacent pairs of nodes gives the following two sets of equations which must both be satisfied in the pairs of indices, there being $2n$ pairs in each set:

$$\begin{aligned} i,i+k &= i+1,\ i+1+2n-k \\ &= i+2,\ i+2+k \\ &= \ldots\ \ldots \\ &= i+j,\ i+j+2n-k \end{aligned} \qquad (4.a)$$

$$= i+j+1,\ i+j+1+k$$

$$= \ldots\ \ldots$$

$$= i+2n-1,\ i+2n-1+2n-k \qquad (4.a)$$

and

$$i, i+2n-k = i+1, i+1+k$$

$$= i+2, i+2+k$$

$$= \ldots\ \ldots$$

$$= i+j, i+j+k \qquad (4.b)$$

$$= i+j+1, i+j+1+2n-k$$

$$= \ldots\ \ldots$$

$$= i+2n-1,\ i+2n-1+k$$

where, in both sets,

$$0 \leqq j \leqq (2n-1) \qquad (4.c)$$

For fixed values of i and k, (4.a) and (4.b) may be written in closed set form as

$$(4.a) = \{i+j, i+j+k \mid j \text{ even, including zero}\}$$

$$\bigcup \{i+j, i+j+2n-k \mid j \text{ nonzero and odd}\}$$

$$(4.b) = \{i+j, i+j+k \mid j \text{ nonzero and odd}\}$$

$$\bigcup \{i+j, i+j+2n-k \mid j \text{ even, including zero}\} \qquad (5)$$

Equation (5) makes it easy to see the alternation in form of the pairs of indices in each set, with even and odd values of j.

The elements in the positions given by (4.a) are identical with each other, as also are the elements in the positions given by (4.b). But those in (4.a) are not necessarily identical with those in (4.b). The next step, therefore, is to determine whether, and under what conditions, duplicity between the two sets can occur. This will depend on k, since, as k varies over its range of values, all elements and their locations in the network are taken into account.

First, when k is equal to n, it is easy to see that the two sets are identical. Further, each set reduces to two successive identical cycles of length n, each with the n distinct pairs of indices,

$$i,i+n;\ i+1,i+1+n;\ \dots\ \dots\ ;\ i+n-1,\ i+2n-1 \tag{6}$$

Next consider the set (4.a), with k different from n. Duplication of index pairs will occur if

either $i+j_1+2n-k = i+j_2$

where j_1 is a nonzero, odd value of j and j_2 is even or zero, provided that, simultaneously (7.a)

$i+j_1 = i+j_2+k$

or $i+j_1+k = i+j_2$

where both j_1 and j_2 are even or zero values of j; provided that, simultaneously (7.b)

$i+j_1 = i+j_2+k$

Case (7.a) is valid for all values of k given by

$$k = 2n - (j_2 - j_1) \tag{8}$$

for then,

$$i+j_2+k = i+j_1+2n = i+j_1$$

Since j_1 is nonzero and odd and j_2 is even or zero, and $2n$ is even, then k in (8) must be odd. Further, from (4.c)

$$i \leqq | j_2 - j_1 | \leqq (2n-1)$$

i.e., the magnitude of (j_2-j_1) is less than the number of terms, $2n$, in (4.a). Which means, since (j_2-j_1) is also nonzero and odd, that each even term in (4.a) is duplicated by an odd term, and vice versa.

Therefore, whenever k is odd, the set (4.a) reduces to two identical cycles each of length n, one of these being the even terms and the other the odd terms, each with the same n distinct pairs of indices. And whenever k is even, the cycle (4.a) is of length $2n$, all $2n$ pairs of indices being distinct.

Similar reasoning applied to (7.b) shows that this case holds only when k is equal to n; which has already been treated as a separate case.

Applying the same type of reasoning to the set (4.b) leads to exactly the same conclusions as for the set (4.a). Still further, that (4.a) and (4.b) are identical sets whenever k is even; and that they have no common pairs of indices when k is odd.

Without any loss of generality, the situation is therefore summarized by selecting all the even terms from (4.a) and all the odd terms from (4.b) as follows:

$$\text{Let } \{A\} = \{i+j, i+j+k \mid j \text{ even, including zero}\}$$

$$\{B\} = \{i+j, i+j+k \mid j \text{ nonzero and odd}\} \qquad (9)$$

for each value of k in the range

$$1 \leqq k \leqq (2n-1)$$

and each value of i from 1 to $2n$, inclusive, with j varying independently over the range

$$0 \leqq j \leqq (2n-1)$$

Whence, in a complete, symmetric n-port, the elements in the positions corresponding to the node-pair indices $\{A\}$ are identical, as also are those in the position given by $\{B\}$. For all even values of k, as well as when k is equal to n (whether n is even or odd), the elements in $\{A\}$ are identical with those in $\{B\}$. For all odd values of k, other than k equal to n with n odd, the elements in $\{A\}$ are different from those in $\{B\}$.

Now consider the union of sets $\{A\}$ and $\{B\}$; i.e., consider

$$\begin{aligned} \{C\} &= \{A\} \cup \{B\} \\ &= \{i+j, i+j+k \mid 0 \leq j \leq 2n-1\} \qquad (10) \end{aligned}$$

On putting $i = 1$ in (10), the set $\{C\}$ becomes identical with the cycle of node pairs given by equation (2), which in turn relates to the symmetric family of two-diagonal positions in the connection matrix of the n-port as illustrated earlier in Figure 3.

With this observation, and using the rules just developed in determining the relation between k and the elements in each cycle, the connection matrix of a complete, symmetric n-port of any order can be written down immediately.

This, then, is a simple and systematic solution to the problem of generating such n-ports, for any value of n.

By way of example, the connection matrix of a complete, symmetric 5-port is given in Figure 6. Figure 7 shows the symmetric sub-networks corresponding to each of the seven different elements in the 5-port; the complete network is the superposition of these seven subnetworks.

IV. SOME COMBINATORIAL PROPERTIES

The number of different kinds of elements in the most general, complete, symmetric n-port may be of interest, and is readily calculated. Suppose this number is E. And note that there are a total of n distinct cycles of the form given by equation (10), in the connection matrix; $(n-1)$ of them being of length $2n$, corresponding to values of k not equal to n, and one cycle of length n, for k equal to n. The total number of different kinds of elements is

$$E = 2N_o + N_e + 1 \tag{11}$$

where N_o, N_e are the number of symmetry cycles in the connection matrix for odd and even values of k, respectively, and the addition of unity accounts for the cycle k equal to n.

It is a strightforward computation to show that, when n is odd,

$$N_o = N_e = (n-1)/2 \tag{12.a}$$

and that, when n is even

$$N_o = (n-2)/2 + 1 \tag{12.b}$$

$$N_e = (n-2)/2$$

Consequently,

$$E = (3n-1)/2 \text{ when } n \text{ is odd}$$

$$E = \frac{3}{2}n \text{ when } n \text{ is even} \tag{13}$$

Another item of interest, in passing, is the total possible number of symmetric subnetworks obtainable from the complete network. Denoting this number by M, it is given by the sum of all possible combinations of E, taken 1, 2, ... , E at a time; for there is a distinct, symmetric subnetwork corresponding to each of the E elements. Consequently,

$$M = \sum_{r=1}^{E} C\,(E,R) \tag{14}$$

$$= (2^E - 1\,)$$

Of course, this includes degenerate and disjoint combinations which, generally, are of little interest or import. For example, subnetworks a and b in Figure 7 are of little consequence on their own, or, for that matter, when combined into a single network. What is of more significance is that M in (14) increases exponentially with E, this indicating the combinatorial complexity of the situation for large n.

V. CONCLUDING COMMENTS

As was stated in the Introduction, the complete, symmetric, two-port lattice has played a most significant role in the historical development of network theory. Hopefully, the complete, symmetric n-port has the potential to contribute similarly in n-port theory; for it represents a logical, topological n-port generalization of the lattice.

There are some obvious further developments which could be undertaken with the symmetric n-port structure; for example, determination of its two-port immittance function, in closed formulas in terms of the elements in the network. This could lead, in turn, to interesting n-port reduction formulas and unbalancing methods, after the manner of lattice reduction and Bartlett's bisection technique in the two-port case.

VI. BIBLIOGRAPHY

1. "The realization of n-port networks without transformers - a panel discussion", IRE Trans. on Circuit Theory, CT-9, 3, September, 1962, pp. 202-214.

2. N. Nakagawa, "On the evaluation of graph trees and driving point admittance", IRE Trans. on Circuit Theory, CT-5, 2, June, 1958, pp. 122-127.

3. John Riordan, An introduction to combinatorial analysis, John Wiley & Sons, Inc., New York, New York, 1958, p. 144.

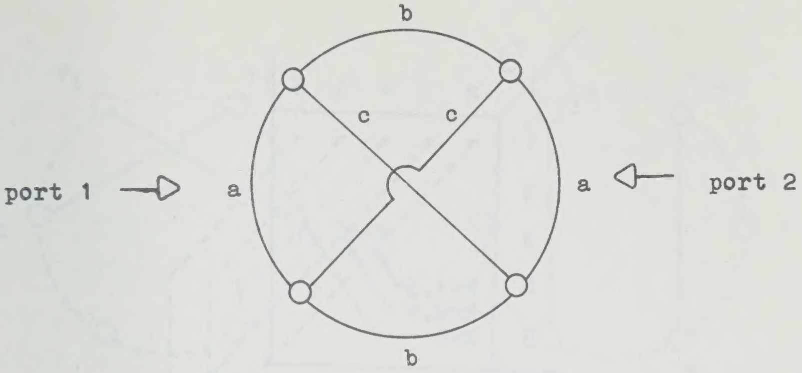
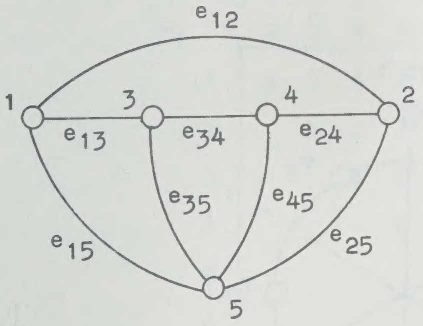

port 1
port 2
a
a
b
b
c
c

Figure 1. Complete, symmetric two-port lattice.

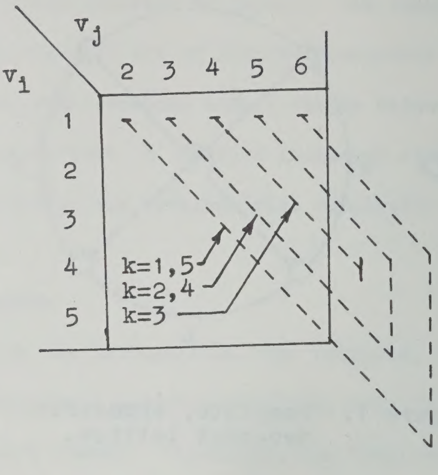

e_{12}
1
3
4
2
e_{13}
e_{34}
e_{24}
e_{35}
e_{45}
e_{15}
e_{25}
5

(a)

$$\begin{array}{c|cccc}
v_i \backslash v_j & 2 & 3 & 4 & 5 \\
\hline
1 & e_{12} & e_{13} & 0 & e_{15} \\
2 & 0 & 0 & e_{24} & e_{25} \\
3 & 0 & 0 & e_{34} & e_{35} \\
4 & 0 & 0 & 0 & e_{45}
\end{array}$$

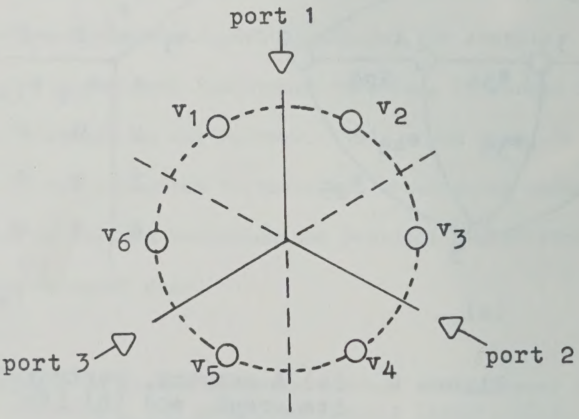
(b)

Figure 2. (a) A network, portrayed by its graph, and (b) its connection matrix.



v_j
v_i
2 3 4 5 6
1
2
3
4
5
k=1,5
k=2,4
k=3

Figure 3. Node-pair cycle symmetries in a 6-node connection matrix.



port 1
v_1
v_2
v_6
v_3
port 3
v_5
v_4
port 2

Figure 4. Symmetry radii in a 3-port.

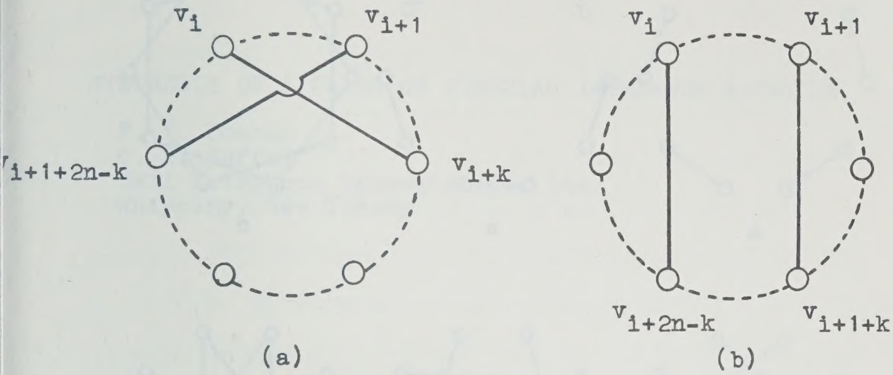

v_i
v_{i+1}
$v_{i+1+2n-k}$
v_{i+k}
(a)
v_i
v_{i+1}
v_{i+2n-k}
v_{i+1+k}
(b)

Figure 5. Element symmetries in a symmetric n-port.

v_i \ v_j	2	3	4	5	6	7	8	9	10
1	a	c	d	f	g	f	e	c	b
2		b	c	e	f	g	f	d	c
3			a	c	d	f	g	f	e
4				b	c	e	f	g	f
5					a	c	d	f	g
6						b	c	e	f
7							a	c	d
8								b	c
9									a

Figure 6. The connection matrix of a complete, symmetric 5-port.

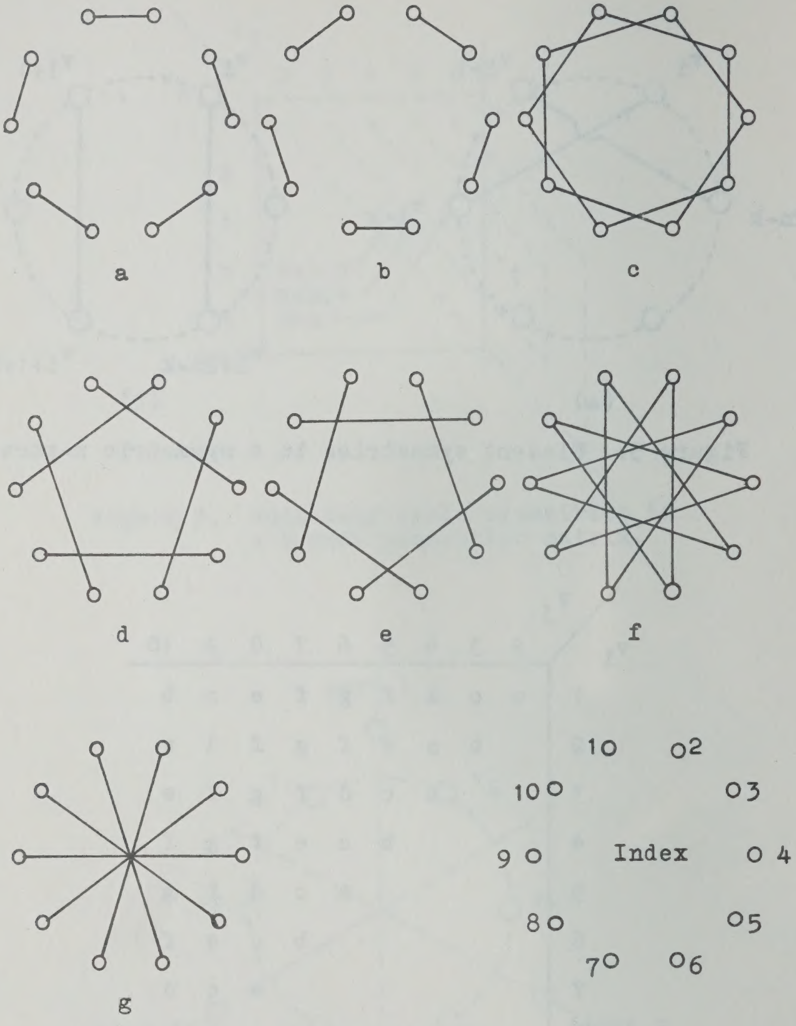

a
b
c
d
e
f
g
1
2
10
3
9
Index
4
8
5
7
6

Figure 7. The symmetric subnetworks corresponding to the different elements in a complete, symmetric 5-port.

SYNTHESIS OF A CLASS OF SINGULAR IMPEDANCE MATRICES

F. T. Boesch
C. L. Suffel
Bell Telephone Laboratories, Inc.
Whippany, New Jersey

SUMMARY

The majority of effort in solving the problem of synthesis of transformerless resistor n-ports has been directed toward the short-circuit admittance matrix. These techniques may certainly be applied to the synthesis of nonsingular open-circuit impedance matrices. However, if the impedance matrix is singular, special procedures must be invoked.

It is easily verified that an impedance matrix is singular if and only if the graph of port voltages contains circuits. Nambiar[1] has given a complete synthesis procedure for the case of all the ports forming a single circuit. It will be shown that the Nambiar result may be generalized to the case of the port voltages forming an arbitrary connected graph. The special case of a port voltage graph which is a tree together with only one link is discussed in detail.

This case, which is the immediate generalization of Nambiar's case, is shown to have similar properties. Possible generalizations to arbitrary singular matrices are also indicated.

INTRODUCTION

The problem of obtaining realizability criteria and synthesis techniques for transformerless resistor n-ports remains a classic problem in network theory. In the last few years, a number of synthesis procedures have been obtained for certain classes of admittance matrices. Now the synthesis of a nonsingular impedance matrix may be accomplished by realizing its inverse as an admittance matrix. However, if the impedance matrix is singular, special procedures are required.

The first synthesis procedure for singular impedance matrices was given by Nambiar(1) who considered impedance matrices realizable on a port structure which is an oriented circuit. It will be shown here that the Nambiar procedure may be generalized to the synthesis of impedance matrices on port structures which are arbitrary connected graphs.

In order to formulate the synthesis procedure, it is necessary to derive some properties of singular impedance matrices. Now in the case of singular impedance matrices, the standard pivotal condensation technique(2) will not yield the impedance matrix. A convenient formulation for the impedance matrix, however, has been given by Bryant.(3) Consider an m terminal network composed entirely of resistors, and let an n-port be defined by means of attaching leads. Let the graph corresponding to the network of resistors be denoted as the network graph, and let the graph obtained by adding arcs corresponding to the ports be called the augmented graph. If $\hat{A}$ denotes the reduced node-arc incidence matrix of the augmented graph, then an appropriate numbering of the arcs enables $\hat{A}$ to be represented as

$$\hat{A} = [SA] \tag{1}$$

where S is the reduced incidence matrix of the network graph, and A is the reduced incidence matrix of the port voltage graph. Let H denote the nodal admittance matrix of the network graph, i.e., H is defined as SGS', where G is the diagonal branch conductance matrix. If H^{-1} exists, then it is easily verified that the open-circuit impedance matrix Z is given by(3)

$$Z = A'H^{-1}A \tag{2}$$

Now it is important to observe that the assumption of the nonsingularity of H does not represent any restriction. This is easily demonstrated by observing that the augmented graph may be assumed to be connected without loss of generality. Furthermore, the definition of an impedance matrix implies that any set of currents may be applied at the ports. Thus if Z exists, the ports cannot form a cut* of the augmented graph. Hence the network graph must be connected. However, it is well known that H is nonsingular if and only if the network graph is connected. Thus the important aspect of the Bryant formula is that it yields Z whenever Z exists. Furthermore a number of properties of Z follow immediately from the Bryant formula. A few of these properties are contained in the next theorem.

Theorem 1

a) rank Z = rank A

b) If A corresponds to a connected graph then

number of nodes = 1 + rank Z

cyclomatic number of A = n - rank Z

c) rank Z = n if and only if the port graph contains no circuits.

Proof

The proof of part - a follows immediately from the following theorem on positive definite matrices:(4)

If P is positive definite and A an arbitrary rectangular matrix then A'PA is positive semi-definite and of rank equal to the rank of A.

Parts - b and c follow immediately from part a.

Now consider the impedance matrix of a network whose port voltage graph is connected but contains circuits. If Z has rank p, then the number of nodes in both the network graph and the port voltage graph is p+1. Thus by open-circuiting n-p ports, a tree port graph is obtained. Hence Z contains a p by p principal minor which is the inverse of an admittance matrix of a p-port on p+1 nodes and utilizes the same network graph as the orignal Z.

---

*In this context, a cut is any set of arcs whose removal disconnects a connected graph.

It is also necessary to review a few of the results of Nambiar.(1) The following definitions are convenient for this purpose.

Definition 1 A symmetric matrix M is said to be indefinite if the column vector consisting of all unit elements is in the null space of M, i.e., if

$$M \underset{\sim}{u} = \underset{\sim}{0} \tag{3}$$

where

$$\underset{\sim}{u} = (1,1,1,\ldots,1)' \tag{4}$$

Definition 2 A symmetric matrix M is said to be potentially indefinite if there exists a column vector of +1 and -1 elements which is in the null space of M, i.e., if there exists a column vector $\underset{\sim}{v}$ of +1 and -1 elements such that

$$M \underset{\sim}{v} = \underset{\sim}{0} \tag{5}$$

Theorem 2(1)

If Z is an n-port impedance matrix whose port graph is a single circuit, then Z is potentially indefinite and has rank n-1. Furthermore, if Z is a realizable open-circuit impedance matrix which is potentially indefinite, then the port graph must be a disjoint union of circuits; if Z has rank n-1 then the port graph is a single circuit.

It is interesting to note that a very simple proof of theorem 2 may be given by utilizing the Bryant formula and the fact that every node is of degree two in a circuit. In fact, this method yields a stronger result which is stated without proof in theorem 3.

Theorem 3

If Z is an n-port impedance matrix with ports $i_1, i_2, i_3, \ldots, i_k$ forming an oriented circuit, then columns $i_1, i_2, i_3, \ldots, i_k$ of Z sum to zero. It is obvious that theorem 3 may also be extended to non-oriented circuits.

It should be observed that in order to determine whether or not a matrix M of rank one less than its order is potentially indefinite it is not necessary to test all possible $\underset{\sim}{v}$ vectors. Notice that the null space of M has dimension one; thus, having obtained any element of the null space, it is immediately obvious that the null space does or does not contain a $\underset{\sim}{v}$ vector.

SYNTHESIS OF SINGULAR IMPEDANCE MATRICES

The properties of singular impedance matrices given in the preceeding section may be utilized as the basis of a synthesis procedure. Consider a real, constant, symmetric matrix Z of order n and rank p. A procedure will be developed which will determine if Z is the open circuit impedance matrix of a transformerless network; if Z is realizable the procedure will yield the network. The first step is to determine a principal minor which corresponds to a tree. In this context it is interesting to observe that the symmetry of Z guarantees the existence of a principal critical minor.(5) Let Z_{pp} denote any principal critical submatrix. Certainly Z_{pp}^{-1} must be realizable as the admittance matrix of a p-port an p+1 nodes. This realization yields a tree of the port graph of Z and the entire network graph appropriate to Z. It merely remains to determine the links of this tree. One method of determining the position of these links is described here. Form all (p+1) by (p+1) principal submatrices $Z_{(p+1)(p+1)}$ of Z which contain Z_{pp}. Each of these $Z_{(p+1)(p+1)}$ submatrices must contain a unique potentially indefinite submatrix of rank one less than its order. In accomplishing a synthesis, however, it is not necessary to test the uniqueness of the potentially indefinite submatrix. The position of the link is determined by the tree ports which are utilized in forming the indefinite matrices. Thus the entire port graph may be determined in this fashion. An alternative method of determining these links is to utilize theorem 3 and consider the entire columns of Z defined by each $Z_{(p+1)(p+1)}$.

In order to verify that this is indeed a realization technique, it must be shown that the open-circuit impedance matrix of the network obtained by this process is equal to the given Z matrix. This is required since certain elements of Z are not utilized in the synthesis; for example, certain link to link mutual impedances never enter into the synthesis procedure. The proof of this last step is quite lengthy and will not be included. However, the proof for the special case of rank n-1 matrices will be given in the next section. Notice that this problem could be avoided if the open-circuit impedance matrix of the constructed network is calculated. If this is not equal to the given Z, then Z is not realizable.

The steps of the synthesis procedure are reviewed in outline form below:

1. Given an $n \times n$ real, constant, symmetric matrix Z of rank p. Obtain a $p \times p$ principal submatrix Z_{pp}.
2. Realize Z_{pp}^{-1} as the admittance matrix of a p-port on p+1 nodes. This yields a tree of the port graph and the entire network graph.
3. Form all (p+1) by (p+1) principal submatrices $Z_{(p+1)(p+1)}$ which contain Z_{pp}. Each of the $Z_{(p+1)(p+1)}$ matrices must contain a potentially indefinite minor of rank one less than its order.
4. The position of the links is determined by the tree ports utilized in forming the potentially indefinite minors.

RANK n-1 IMPEDANCE MATRICES

The case which is the immediate generalization of Nambiar's case is the rank n-1 impedance matrix. In this case, the port structure is a tree together with one link; Nambiar's case is a linear tree whose tip nodes are joined by a link. In his case, a principal submatrix of order n-1 is obtained and realized on a linear tree. The fact that the open circuit impedance matrix of the derived network is equal to the given Z matrix follows from the potentially indefinite nature of Z. More precisely, the open circuit impedance matrix and Z are both potentially indefinite with respect to the same $\underset{\sim}{v}$ vector. Thus it follows from the equality of their principal minors of order n-1 that they are equal.

It is shown in this section that a similar property may be obtained in the general rank n-1 case. The approach is to prove that the entire Z matrix is uniquely determined by the properties of Z which enable the synthesis of the derived network via the procedure given in the preceeding section. In fact, the result is that the uniqueness is guaranteed by the positive definiteness of $Z_{(n-1)(n-1)}$ and the existence of a potentially indefinite principal submatrix. Notice that the realizability of the inverse of $Z_{(n-1)(n-1)}$ certainly implies that it is positive definite.

In order to accomplish the proof, it is convenient to utilize the following lemma.

Lemma

Consider the matrix P defined as

$$P = \begin{bmatrix} L & \underset{\sim}{x} \\ \underset{\sim}{x}' & 0 \end{bmatrix} \tag{6}$$

where L is a real, positive definite, symmetric matrix and $\underset{\sim}{x}$ a real column vector. Then P is singular if and only if $\underset{\sim}{x}$ is the zero vector.

Proof

If x is the zero vector, then the result is obvious. Now suppose that P is singular, then there exists a non-zero column vector $\underset{\sim}{\omega}$ such that

$$P \underset{\sim}{\omega} = 0 \tag{7}$$

Let $\underset{\sim}{\omega}$ be partitioned conformally with P into $\underset{\sim}{\omega}_1$ and $\underset{\sim}{\omega}_2$, where ω_2 is a scalar. Thus

$$L \underset{\sim}{\omega}_1 + \underset{\sim}{\omega}_2 \underset{\sim}{x} = 0 \tag{8}$$

and

$$\underset{\sim}{x}' \underset{\sim}{\omega}_1 = 0 \tag{9}$$

Now multiply equation 8 on the left by $\underset{\sim}{\omega}_1'$ and utilize the ninth equation to obtain

$$\underset{\sim}{\omega}_1' L \underset{\sim}{\omega}_1 = 0 \tag{10}$$

Since L is positive definite,

$$\underset{\sim}{\omega}_1 = 0 \tag{11}$$

Thus the equation

$$\omega_2 \underset{\sim}{x} = 0 \tag{12}$$

implies that the only possibility for the existence of a non-zero $\underset{\sim}{\omega}$ vector is that $\underset{\sim}{x}$ be a zero vector. Q.E.D.

Now in the case of rank n-1 Z matrices, let

$$Z_{(n-1)(n-1)} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z'_{12} & Z_{22} \end{bmatrix} \tag{13}$$

where the partition is accomplished such that Z_{22} together with the corresponding elements of the n^{th} row and column of Z form a potentially indefinite matrix. For convenience, assume that the signs of Z have been changed such that this potentially indefinite matrix becomes indefinite. Now let $\underset{\sim}{u}$ denote a column vector all of whose elements are unity. Thus the original Z matrix may be written as

$$Z = \begin{bmatrix} Z_{11} & Z_{12} & Z_{13} \\ Z'_{12} & Z_{22} & -Z_{22}\underset{\sim}{u} \\ Z'_{13} & -\underset{\sim}{u}'Z_{22} & \underset{\sim}{u}'Z_{22}\underset{\sim}{u} \end{bmatrix} \tag{14}$$

where Z_{13} is of course a column vector. The next theorem shows that Z_{13} is uniquely determined.

Theorem 4

Consider an arbitrary real symmetric matrix Z of the form given in equation 14, such that $Z_{(n-1)(n-1)}$ is positive definite. Then Z is singular if and only if

$$Z_{13} = -Z_{12}\underset{\sim}{u} \tag{15}$$

Proof

If

$$Z_{13} = -Z_{12}\underset{\sim}{u} \tag{16}$$

the result is obviously singular. Now suppose Z is singular. Then the matrix L given by

$$L = \begin{bmatrix} Z_{11} & Z_{12} & Z_{13} + Z_{12}\underset{\sim}{u} \\ Z_{12}' & Z_{22} & 0 \\ Z_{13}' + \underset{\sim}{u}'Z_{12}' & 0 & 0 \end{bmatrix} \tag{17}$$

is obtained by elementary operations on Z and is therefore also singular. Now by invoking the lemma,

$$Z_{13} = -Z_{12}\underset{\sim}{u} \tag{18}$$

Thus the entire synthesis procedure is validated in the rank n-1 case. The general case of an arbitrary rank may be treated by a repeated application of theorem 4 to certain submatrices of Z. Unfortunately, the details involved in the description of this process are much to lengthy to be included here.

CONCLUSIONS

It has been shown that the Bryant formula provides a convenient basis for obtaining properties of impedance matrices of resistor n-ports. In particular it may be used to analyze singular impedance matrices. It should be noted that the standard method of augmentataion and pivotal condensation is not valid in the singular case. It is indeed interesting to note that the derivation of paramount[2] as a necessary condition is not valid for the singular case since it is based on pivotal condensation.

It has been shown that the synthesis given by Nambiar in his original work may be extended to arbitrary connected port graphs. Furthermore the only reason for restricting the procedure to connected graphs is that known results may be utilized to synthesize an admittance matrix on a tree of port voltages. If the synthesis of admittance matrices can be generalized to the case of a disjoint collection of trees, then the procedure given here for singular impedance matrices could easily be generalized to arbitrary port graphs. Thus it can be assumed that the synthesis of admittance matrices is the only case that need be treated.

Finally, it should be mentioned that, after the completion of this work, the authors became aware of the most recent results of Nambiar.(6) In this paper

Nambiar discusses the same problem. However, the methods given are different and are restricted to non-separable port graphs.

REFERENCES

1. K. K. Nambiar, "Answer to Foster's Open Question", IEEE Trans. on Ckt. Th., Vol. CT-10, March 1963, p. 127.

2. I. Cederbaum, "Conditions for Impedance and Admittance Matrices of n-Ports Without Ideal Transformers", Proc. IEE, Part c, Mono. 276R, p. 245, January 1958.

3. P. Bryant, "Discussion on Conditions for Impedance and Admittance Matrices of n-Ports Without Ideal Transformers", Proc. IEE, Vol. 106, part c, p. 116, 1959.

4. S. N. Roy, Some Aspects of Multivariate Analysis, Wiley, p. 142.

5. C. C. MacDuffee, The Theory of Matrices, Chelsea, p. 12.

6. K. K. Nambiar, "On the Realization of Singular R Matrices", IEEE Trans. on Ckt. Th., Vol. CT-11, September 1964, p. 421.

DIGITAL COMPUTER ANALYSIS OF PASSIVE NETWORKS USING TOPOLOGICAL FORMULA

George W. Zobrist and Gladwyn V. Lago
Department of Electrical Engineering
University of Missouri
Columbia, Missouri

I. Introduction

In a classic paper by Kirchhoff(1) the rules for computation of network response (in terms of branch resistances) by means of topological quantities were presented. About 40 years later Maxwell(2) presented similar rules, dual to those of Kirchhoff, in terms of conductances. These rules will be referred to as Kirchhoff's rules. Interest has been generated in the last few years with applications to synthesis(3-5) and investigations of the applicability of these topological rules for analysis of networks using the digital computer have been presented by Mayeda and Van Valkenburg,(6) Hobbs,(7) and MacWilliams.(8)

The intent of this paper is to present a general approach whereby all trees for a v vertex complete graph are generated and any subset of trees for a subgraph of the complete graph of n vertices, where $n \leq v$, can be determined. The network response, through use of Kirchhoff's rules, can be determined for any linear passive network (with no mutual inductances) once the trees are found for the networks graph.

In conjunction with this aim two problems had to be solved: First a method of obtaining the trees for a m vertex complete graph from a v vertex complete graph. The trees for a subgraph of m vertices can be readily found from a complete graph of m vertices. Secondly, a least upper bound on the number of 2-trees (2-cotrees). This is required for computer storage reasons, since one has to compare 2-tree (2-cotrees) for determination of the sign of each 2-tree (2-cotree) product, in transfer quantities.

Computer programs were written employing these ideas and checked for graphs with up to six vertices. In principle the number of vertices allowed is limited only by storage capacity of the computer and time. Fortran II-D language was chosen since it seems to be in rather wide spread use.

II. Kirchhoff's Topological Rules

The main purpose of this section will be to state Kirchhoff's topological rules in a notation which was found rather useful in the formulation of the computer programs. The reader is referenced to excellent discussions in either Seshu(9) or Weinberg(10) for detailed developments. Definitions of terms used are the same as found in Seshu.(9)

While the original rules were formulated only for resistances (conductances) they are easily extended to the R-L-C case;(10) in the remainder of the paper this will be assumed.

If a current source is attached parallel to a branch which includes an element y_m (see Figure 1(a)), the voltage response, at this pair of terminals, is given by

$$V(s) = Ig(s) \frac{\sum_{k=1}^{\text{Trees}} \frac{\partial}{\partial y_m} \left(y_{k1} \cdots y_{k(v-1)} \right)}{\sum_{G=1}^{T} \left(y_{G1} \cdots y_{G(v-1)} \right)} ,$$

where the denominator is the sum of all tree products for the given networks graph of e edges and v vertices and the numerator is the sum of all 2-tree products, with respect to y_m. The notation $y_{k1} \cdots y_{k(v-1)}$ refers to the tree product of the k^{th} tree, which consists of the v-1 distinct edges of the graph that form the k^{th} tree. Where y_{k1} is not necessarily equal to y_{gi}, if $g \neq k$. If y_m or y_n occur in a tree product, the $\frac{\partial}{\partial y_m}$ or $\frac{\partial}{\partial y_n}$ of $(y_{k1} \cdots y_{k(v-1)}) \neq 0$, if not, it is zero.

If a current source is attached to a pair of terminals (see Figure 1(b)), the voltage response at a pair of terminals is given by

$$V(s)=Ig(s) \frac{\sum_{k=1}^{T} \sum_{f=1}^{T} \pm \left(\frac{\partial}{\partial y_m} (y_{k1} \cdots y_{k(v-1)}) \cap \frac{\partial}{\partial y_n} (y_{k1} \cdots y_{k(v-1)}) \right)}{\sum_{G=1}^{T} (y_{G1} \cdots y_{G(v-1)})} .$$

The denominator is interpreted as before but the numerator takes on different significance. The numerator represents the algebraic sum of all common 2-trees, where the 2-trees are originally separated into two sets, one consisting of all tree products containing y_m, the other set those tree products containing y_n. The common terms in these two sets are the desired 2-tree products.

The + sign is used if upon shorting of the edges of the above mentioned 2-tree product one finds that the current source will cause the assigned positive voltage drop, negative if a negative voltage drop is produced.

If a voltage source is placed in series with a branch which contains z_m (see Figure 1(c)), the current response in this branch is given by

$$I(s)=Vg(s)\;\frac{\sum_{D=1}^{T}\frac{\partial}{\partial Z_m}(Z_{D1}\cdots Z_{DL})}{\sum_{H=1}^{T}(Z_{H1}\cdots Z_{HL})},$$

where L = e - (v-1). The notation $z_{k1}\ \dots\ z_{kL}$ is similar to that of the admittance equations, except now it is the k^{th} cotree and L distinct edges.

The denominator is the sum of cotree products, for a given graph of e edges and v vertices and the numerator is the sum of all 2-cotree products, with respect to z_m.

If a voltage source is placed in series with a branch which contains z_m (see Figure 1(d)), the current response in a branch which contains z_n is given by,

$$I(s)=Vg(s)\;\frac{\sum_{D=1}^{T}\sum_{C=1}^{T}\pm\left(\frac{\partial}{\partial Z_m}(Z_{D1}\cdots Z_{DL})\cap\frac{\partial}{\partial Z_n}(Z_{C1}\cdots Z_{CL})\right)}{\sum_{H=1}^{T}(Z_{H1}\cdots Z_{HL})}\,.$$

The denominator is interpretated the same as in the driving point impedance case and the numerator consists of the algebraic sum of the common 2-cotree terms. Where the + sign is used, if upon removal of the 2-cotree under investigation from the network, the voltage source causes a current which is confluent with the assigned positive direction, negative if not.

III. Concept of the Complete Graph

It is a well known fact (see Cayley$^{(11)}$), that the number of trees in a complete graph, i.e., a graph with a single edge from every vertex to every other vertex, is v^{v-2}. There are also many established methods of enumerating all the trees of a graph.$^{(12-15)}$ The method of obtaining trees in this paper follows a theorem stated in Seshu,$^{(9)}$ i.e., if G_s is a subgraph of a connected graph G of v vertices, and G_s consists of v-1 edges and no circuits, then G_s is a tree of G. This theorem was found to be easily implemented on the digital computer and since the intent was to generate the complete set of trees only once, time was not of first importance.

It was of prime importance to be able to obtain the trees of a complete graph of m vertices from the trees of a complete graph of v vertices, $m < v$. This can be done in the following manner. Assume a complete graph of v vertices. This complete graph has $v(v-1)/2$ edges. Now remove edge (1,v) and let vertices 1 and v coalesce, the trees of this resultant graph contains all trees (when edge (1,v) is included in each tree of the resultant graph) which contain edge (1,v). Remove edge (1,v) from the complete graph of v vertices. Now remove edge (2,v) and let vertices 2 and v coalesce, the resultant graphs trees contain all trees of the original complete graph, which contain edge (2,v) but not edge

(1,v), again when edge (2,v) is included. Continue in this manner until edge (v-1,v) is reached. This development separates the trees of a complete graph of v vertices into v-1 distinct sets. The first set contains all the trees which have edge (1,v), the second set consists of those trees which do not have edge (1,v), but do have edge (2,v). This process is continued until one reaches the last set, this set contains edge (v-1,v) but not edges (1,v), (2,v), ---, (v-2,v). Collecting all the trees for the v-1 sets, one has obtained the trees of the v vertex complete graph. Of interest is the last set of trees, those which contain edge (v-1,v) but not edges (1,v) --- (v-2,v). This set is found to contain the trees of a complete graph of v-1 vertices, see Figure 2.

This can be shown to be a valid development by considering the nodal determinant of a v vertex complete graph, since the above can be obtained by expanding the determinant in a particular manner. In fact, in this expansion (see Figure 2), each resultant subgraph is a quasi-complete graph, i.e., some vertices may have parallel edges between them. In summary, one may find the trees of a m vertex complete graph from a v vertex complete graph, quite easily, by considering only those trees which contain certain edges. Another interesting point is that this procedure can also be used to find the trees of a complete graph of v + 1 vertices from a complete graph of v vertices.

If one is to store 2-trees (2-cotrees) in an efficient manner, a least upper bound has to be determined for them. It is asserted that the number of 2-trees for a complete graph of v vertices is given by $2\ v^{v-3}$. This can be seen in the following manner. Consider the following expression,

$$\begin{array}{c}\text{Total \# edges}\\ \text{for complete}\\ \text{graph}\end{array} \quad \mathrm{X} \quad \begin{array}{l}\text{Total \# 2-trees for}\\ \text{complete graph for}\\ \text{a given edge (i,j)}\end{array} \quad \mathrm{X} \quad \frac{1}{v-1} \quad .$$

Now due to symmetry the total number of 2-trees for edge (i,j) is the same as edge (k,l), hence if one takes the product of the first two terms one obtains a term which would be a tree, since this is the reverse process of obtaining a 2-tree. But enumerating all these trees would give repetitions since there are v-1 terms in each product, hence each product will occur in v-1 sets of the above total. Therefore, if one multiplies by the factor $\frac{1}{v-1}$, this expression can be equated to the total number of trees in a complete graph of v vertices. Upon substitution for the total number of edges its equivalent in terms of v, one has

$$\frac{v(v-1)}{2} \times \# \text{ 2-trees} \times \frac{1}{v-1} = v^{v-2} \text{ or } \# \text{ 2-trees} = 2\,v^{v-3} .$$

This establishes a least upper bound for the number of 2-trees in a graph of v vertices, with no parallel edges. If parallel edges are to be allowed, a similar development shows that the number of 2-trees for a full graph, i.e., a complete graph with p parallel edges between each vertex, is $2p(pv)^{v-3}$.

The number of 2-cotrees for a complete graph of v vertices and a full graph of v vertices and p parallel edges, is found to be

$$\# \text{ 2-cotrees} \begin{array}{ll} = v^{v-2} - 2v^{v-3} & \text{(complete)} \\ = p(pv)^{v-2} - 2p\,(pv)^{v-3} & \text{(full)} \end{array} .$$

This is easily deduced since if a tree contains edge (i,j) the associated cotree does not and vice versa.

IV. Method of Obtaining Results on Computer

Edges are designated by the unordered pair (v_i, v_j), i.e., the edge connecting vertex 4 to vertex 7 would be (4,7). This designation allows efficient processing on the digital computer.

Trees are generated in the following manner; form all combinations of the v(v-1)/2 edges taken v-1 at a time and test each combination to see if it is a tree or not. This test can be done by removing all free edges, i.e., any edge connected to a vertex of degree 1. After this removal has been done for the original combination, remove the free edges of the resulting combination of edges. Continue this until no free edges remain.

In doing this the original graph can be reduced to all isolated vertices (if edges and vertices are considered to be separate) if it is a tree.

Example of method of finding the trees for a 4 vertex complete graph. The edges are: (1,2)(1,3)(1,4)(2,3)(2,4)(3,4). Now form the combinations: (1,2)(1,3)(1,4); (1,2)(1,3)(2,3); (1,2)(1,3)(2,4); (1,2)(1,3)(3,4); (1,2)(1,4)(2,3); (1,2)(1,4)(2,4); (1,2)(1,4)(3,4); (1,2)(2,3)(2,4); (1,2)(2,3)(3,4); (1,2)(2,4)(3,4); (1,3)(1,4)(2,3); (1,3)(1,4)(2,4); (1,3)(1,4)(3,4); (1,3)(2,3)(2,4); (1,3)(2,3)(3,4); (1,3)(2,4)(3,4); (1,4)(2,3)(2,4); (1,4)(2,3)(3,4); (1,4)(2,4)(3,4). The second, sixth and thirteenth combinations do not form trees. These are all the combinations which need to be formed. The next one is (2,3)(2,4)(3,4) and since each vertex must be included in a tree every combination after the last one listed need not be formed. In general there is a saving of $\binom{\frac{v(v-1)}{2} - (v-1)}{V-1}$ combinations to be tested in forming this way.

If all vertices are not isolated and no more free edges are encountered, the graph contains at least one circuit, since no vertices have degree 1 (see Figure 3), hence the subgraph for this combination does not form a tree.

In Figure 4 is a somewhat simplified flow diagram of the computer program for calculating driving point and transfer admittances.

Trees are stored on the IBM Disk Pack 1311 as floating point numbers to utilize to the maximum available storage space, although in doing this one trades off time in converting back to fixed point for calculation purposes. As each tree is drawn off the disk a check is performed to see if it is a valid tree for the complete graph specified. This is done by checking to see if the tree product contains only the edges (v-1,v), (v-2,v-1), ..., (i-1,i), and no other edges containing v,v-1, ..., i, where i-1 is the number of vertices in the specified complete graph.

The tree product is checked to see if it is zero, for the given subgraph of the complete graph with i-1 vertices, if so, the next tree on the disk is investigated. If the tree product is non-zero the computer proceeds to find out whether it is a 2-tree. If it contains y_m (see Figure 1(b)), set y_m to 1, which is equivalent to performing the operation $\frac{\partial}{\partial y_m}$, since each tree product is linear in any given y. Convert the 2-tree product to floating point and store on the disk. Similarly, if transfer quantities are desired, check for 2-trees containing y_n (see Figure 1(b)); these are also stored on the disk. The necessity for determining a least upper bound on 2-trees now becomes evident (see Figure 5).

Since quantities on the disk cannot be given variable names, and there is no pattern as to when the 2-trees for y_m and y_n will be encountered, a method such as this in Figure 5 alleviates devising complicated recalling routines.

The 2-trees are compared and the common ones found and checked for sign by the following intuitive scheme (see Figure 6) due to Percival.(14)

If at starting at 1 one ends up at 2' rather than 2, a negative sign is associated with the 2-tree product. To keep track of powers of s in the various tree and 2-tree products one makes the following association $s^{-1} \rightarrow -1$, $s^{1} \rightarrow 1$, $s^{0} \rightarrow 0$, where each term in a product is assigned either 1, -1, or 0 and then all terms summed.

The procedure for driving point admittance is simpler since only the 2-trees for y_m are found and hence no comparison or determination of sign is needed.

In Figure 7 is a simplified flow diagram of the computer program for calculating driving point and transfer impedances.

In determining impedance functions one needs to find the cotrees; this can be done by negation. If one removes from the edges of the graph the edges of the non-zero tree product under investigation, the remaining edges form a cotree for the given graph. Then each step is performed as was done for the admittance case. One exception is that one has to remove the 2-cotree from the edges of the graph to check for a + or - sign on common 2-cotrees. This is expected since 2-cotrees can be considered to be the dual of 2-trees.

V. Examples

Y (s) Transfer Admittance

Values of elements are for admittances. Correspondence of (i,j) = k is made for input i.e., (5,6) = 15, (3,4) = 10, etc. Correspondence is made for 6 node complete graph.

Input:

Card # 1

1. Number of vertices = 6

2. Edge which current source is in parallel with = 1

Card #2

1. Edge across which voltage response is to be calculated = 10

2. Vertex which is connected to arrowhead of current source = 1

3. Vertex which is connected to + of V(s) = 4

4. Vertex which is connected to - of V(s) = 3

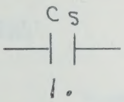
Card #3

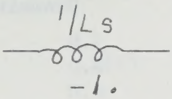
Admittance values of edges of graph of given network

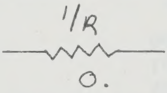
(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)
1.	0.	0.	1.	0.	0.	0.	0.	1.	1.	1.	1.	1.	1.	0.

Card #4

Value of admittance element, according to whether it is a







1. 0. 0. 1. 0. 0. 0. 0. -1. 0. -1. 1. 0. 1. 0.

Output:

```
                              Y (s)

     Numerator                                    Denominator

    10.0      SPO                            -1.0        SPO
     8.0      SP1                             1.0        SP1
     9.0      SP2                             0.0        SP2
     4.0      SP3                             0.0        SP3
     1.0      SP4                             0.0        SP4
     0.0      SP5                             0.0        SM1
```

```
        Numerator                              Denominator

      0.0       SM1                          0.0      SM2
      0.0       SM2                          0.0      SM3
      0.0       SM3                          0.0      SM4
      0.0       SM4
      0.0       SM5
```

Y(s) Driving Point Admittance

Input:

Card #1

1. Number of vertices = 6

2. Edge which current source is in parallel with = 1

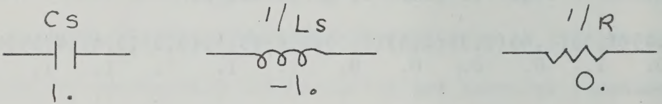
Card #2

Admittance values of edges of graph of given network

1. 0. 0. 1. 0. 0. 0. 0. 1. 1. 1. 1. 1. 1. 0.

Card #3

Value of admittance element, according to whether it is



1. 0. 0. 1. 0. 0. 0. 0. -1. 0. -1. 1. 0. 1. 0.

Output:

```
                              Y (s)

      Numerator                                Denominator

    10.0       SP0                           5.0       SP0
     8.0       SP1                           8.0       SP1
     9.0       SP2                           4.0       SP2
     4.0       SP3                           1.0       SP3
     1.0       SP4                           0.0       SP4
     0.0       SP5                           6.0       SM1
     0.0       SM1                           0.0       SM2
     0.0       SM2                           0.0       SM3
     0.0       SM3                           0.0       SM4
     0.0       SM4
     0.0       SM5
```

Z(s) Transfer Impedance

Values of element are for impedances. Correspondence of (i,j) = k is made for input, i.e., (3,4) = 10, (5,6) = 15, etc.

Correspondence is made for 6 node complete graph.

Input:

Card #1

1. Number of vertices = 4
2. Edge which voltage source is inserted in series with = 1
3. Total number of edges in graph = 6

Card #2

1. Edge through which current response is to be calculated = 10
2. Vertex which is connected to + side of $V_g(s)$ = 1
3. Vertex which is situated at end opposite arrow head of I(s) = 3
4. Vertex which is situated at arrowhead of I(s) = 4

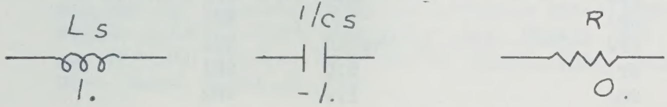
Card #3

Impedance values of edges of graph of given network

```
   1.  2.  1.  0.  0.  1.  1.  0.  0.  3.  0.  0.  0.  0.  0.
```

Card #4

Value of impedance element, according to whether it is a



```
   0.  1.  -1.  0.  0.  1.  -1.  0.  0.  1.  0.  0.  0.  0.  0.
```

Output:

```
                        Z(s)

       Numerator                          Denominator

      9.0     SP0                        -1.0       SP0
     22.0     SP1                         0.0       SP1
     11.0     SP2                         0.0       SP2
      0.0     SP3                         0.0       SM1
      3.0     SM1                         0.0       SM2
      1.0     SM2
      0.0     SM3
```

Z(s) Driving Point

Input:

Card #1

1. Number of vertices = 4

2. Edge which voltage source is inserted in series with = 1

3. Total number of edges in graph = 6

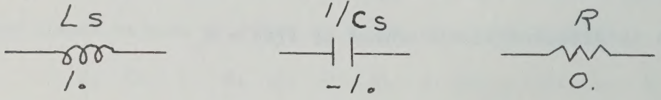
Card #2

Impedance values of edges of graph of given network

1. 2. 1. 0. 0. 1. 1. 0. 0. 3. 0. 0. 0. 0. 0.

Card #3

Value of impedance element, according to whether it is a



0. 1. -1. 0. 0. 1. -1. 0. 0. 1. 0. 0. 0. 0. 0.

Output:

```
                         Z(s)

      Numerator                      Denominator

   9.0       SP0                  9.0       SP0
  22.0       SP1                  0.0       SP1
  11.0       SP2                 11.0       SP2
   0.0       SP3                  0.0       SM1
   3.0       SM1                  1.0       SM2
   1.0       SM2
   0.0       SM3
```

See Figure 8 for network configurations used in examples.

VI. Conclusions

The method is completely general in that the response for any network of n vertices can be found from the knowledge of the trees for a complete graph of v vertices, where n $\leq$ v.

In using this approach the size network (number of vertices) which can be handled is reached rather fast on IBM 1620 with disk pack. Only graphs with up to 7 vertices can be accommodated. On the IBM 1620 it takes approximately one second to perform a complete cycle, i.e., time lapse between drawing off the j^{th} tree to drawing off the $(j+1)^{th}$ tree.

Some rather interesting results have been obtained. They are the establishment of a least upper bound on 2-trees (2-cotrees) and a method of obtaining the trees of a complete graph of m vertices from one of v vertices, where $m < v$. Copies of the computer programs are available.

ACKNOWLEDGMENTS

This paper was partially supported by the National Science Foundation and the University of Missouri Engineering Experiment Station. This paper also represents a portion of the work in progress on a Ph.D. dissertation, by the first author, at the University of Missouri, Columbia, Missouri.

BIBLIOGRAPHY

1. Kirchhoff, G., "Über die Auflösung der Gleichungen, auf welche man bei der Untersuchungen der Linearen Verteilung Galvanisher Ströme geführt wird", Poggendorf Ann. Physik, Vol. 72, (1847) pp. 497-508; English translation, Trans. Inst. Radio Engrs. CT-5 (March 1958) pp. 4-7.

2. Maxwell, J. C., Electricity and Magnetism, Vol. 1, Chap. VI and Appendix. Oxford: Clarendon Press, 1892, pp. 403-410.

3. Seshu, S., "Topological Considerations in the Design of Driving Point Functions", Trans. Inst. Radio Engrs. CT-2 (December, 1955) pp. 356-367.

4. Kim, W. H., "A New Method of Synthesis of Driving Point Functions", Interim Tech. Rpt. No. 1, U.S. Army Contract No. DA-11-022-PRD-1938, U. of Ill.; April 1956.

5. Talbot, A., "Some Fundamental Properties of Networks Without Mutual Inductance", Proc. Inst. Elec. Engrs., Vol. 102 (September, 1955) pp. 168-175.

6. Mayeda, W., and Van Valkenburg, M. E., "Network Analysis and Synthesis by Digital Computer", Inst. Radio Engrs. Convention Record (part 2), Vol. 5 (1957) pp. 137-144.

7. Hobbs, E. W., "Topological Network Analysis as a Computer Program", (Discussion), IRE Trans. PGCT-6 (1959) pp. 135-36.

8. MacWilliams, F. J., "Topological Network Analysis as a Computer Program", IRE Trans. PGCT-5 (1958) pp. 228-29.

9. Seshu, S. and Reed, M. B., Linear Graphs and Electrical Networks. Reading, Mass.: Addison-Wesley, 1961.

10. Weinberg, L., Network Analysis and Synthesis. New York: McGraw-Hill, 1962.

11. Cayley, A., "A Theorem on Trees", Quarterly Journal of Pure and Applied Math, Vol. XXIII (1889) pp. 376-378.

12. Nakagawa, N., "On Evaluation of the Graph Trees and the Driving Point Admittance", IRE Transactions of Circuit Theory CT-5, No. 2 (June 1958) pp. 122-127.

13. Cummins, R. L. and Thomason, L. C., "An Efficient Tree-listing Program", Southwestern IEEE Conference, 1964.

14. Percival, W. S., "Solution of Passive Electrical Networks by Means of Mathematical Trees", J. Inst. Elect. Engrs. (London), Vol. 100, (Part III) (1953) pp. 143-150.

15. Hakimi, S. S., "On Trees of a Graph and Their Generation", Franklin Institute Journal, Vol. 272, No. 5 (1961) pp. 347-359.

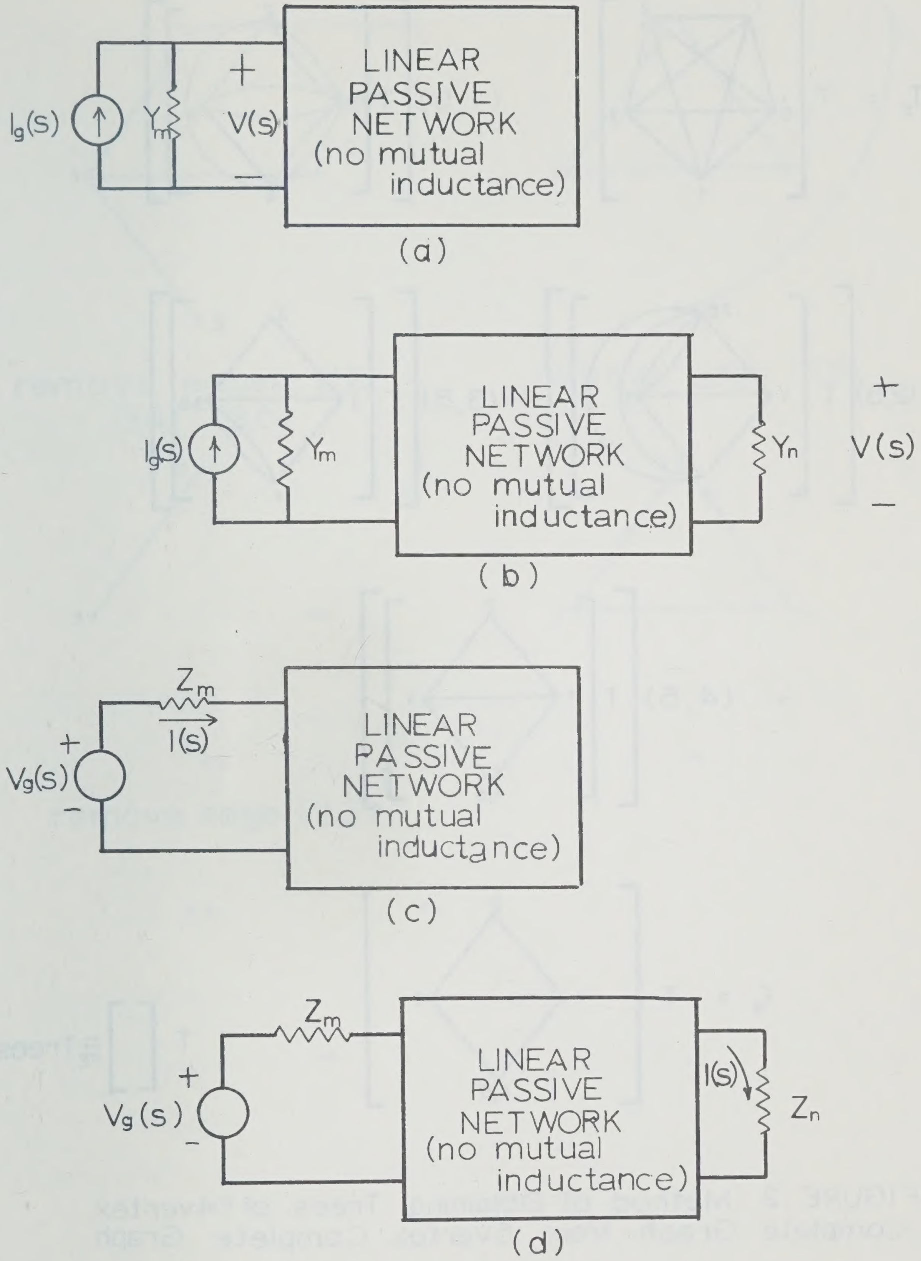

$I_g(s)$
Y_m
+
$V(s)$
−
LINEAR PASSIVE NETWORK (no mutual inductance)
(a)
$I_g(s)$
Y_m
LINEAR PASSIVE NETWORK (no mutual inductance)
Y_n
+
$V(s)$
−
(b)
Z_m
$I(s)$
+
$V_g(s)$
−
LINEAR PASSIVE NETWORK (no mutual inductance)
(c)
Z_m
+
$V_g(s)$
−
LINEAR PASSIVE NETWORK (no mutual inductance)
$I(s)$
Z_n
(d)

FIGURE 1. Network Configurations for Application of Kirchoff's Topolological Rules

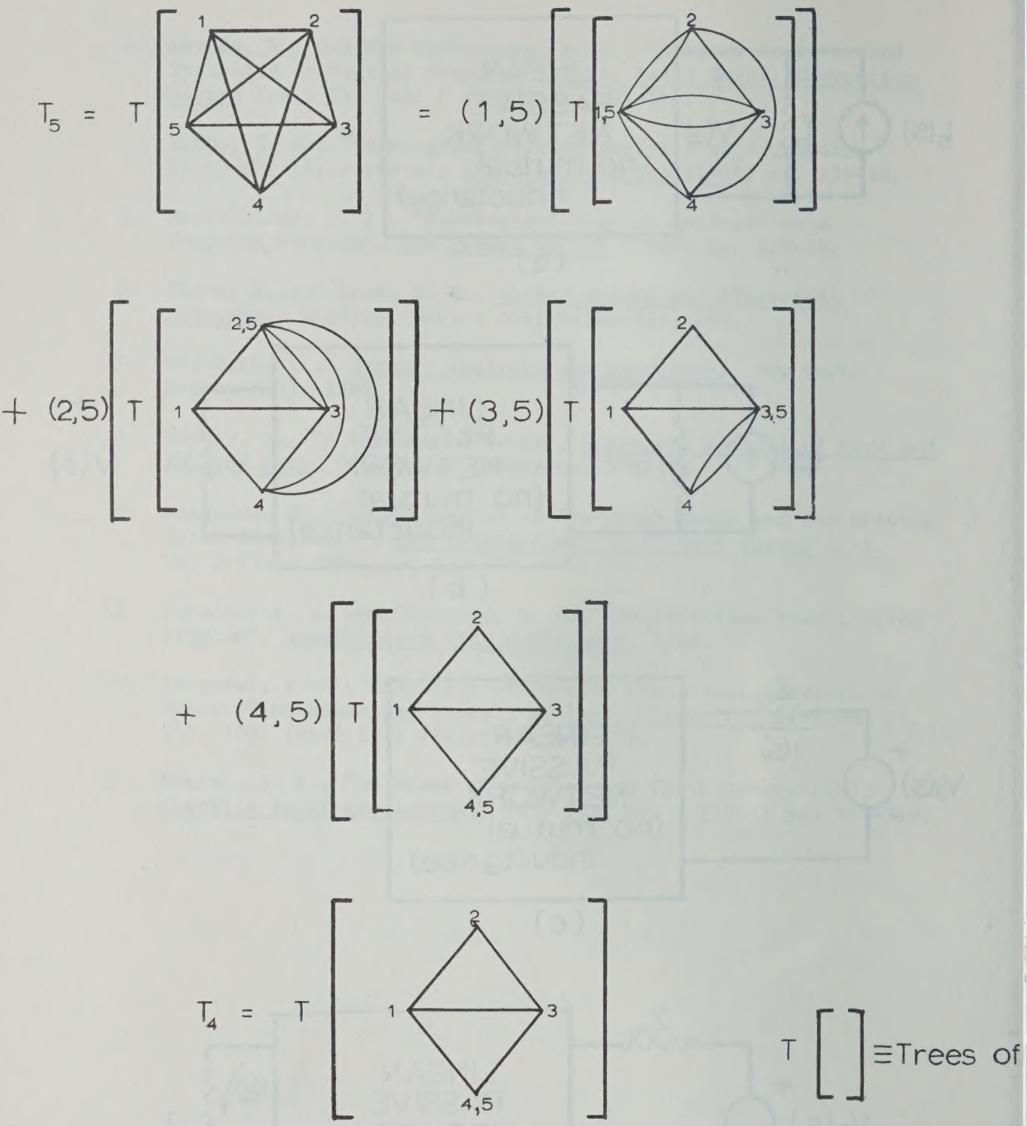

T5 = T
1
2
3
4
5
= (1,5) T
1,5
+ (2,5) T
2,5
+ (3,5) T
3,5
+ (4,5) T
4,5
T4 = T
T [ ] ≡Trees of

FIGURE 2. Method of Obtaining Trees of 4vertex Complete Graph from 5vertex Complete Graph

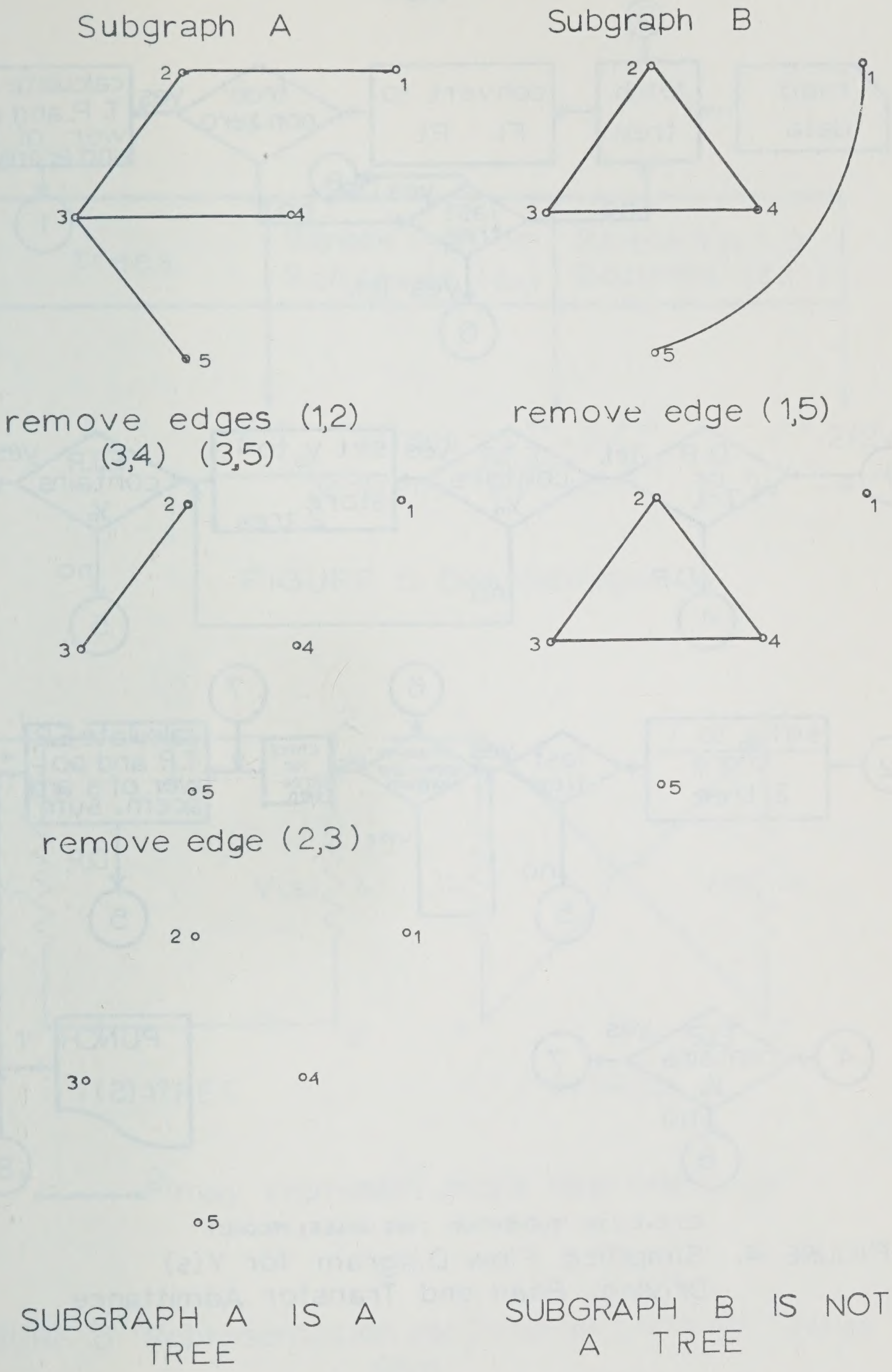
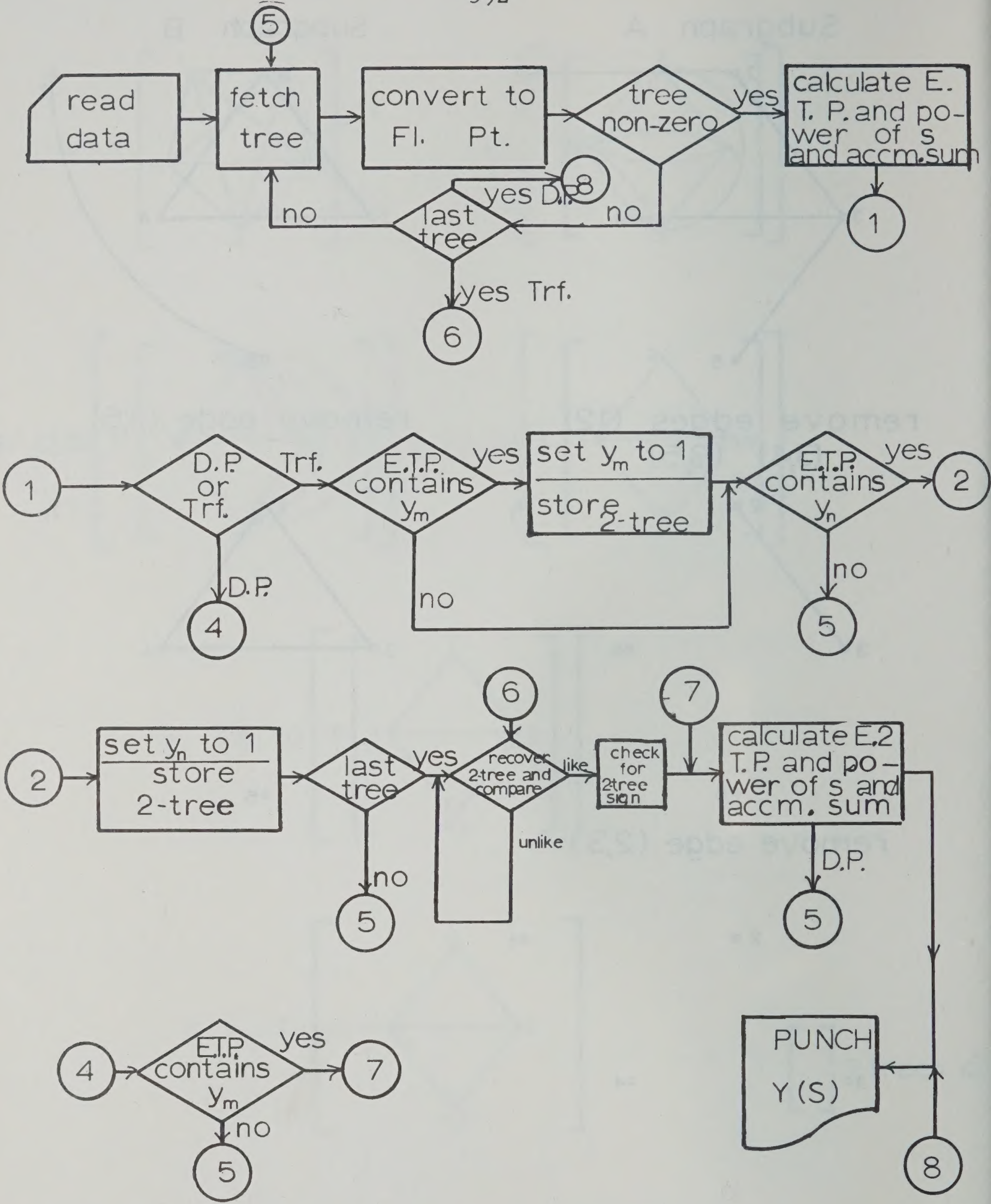

Subgraph A
Subgraph B
2
1
3
4
5
2
1
3
4
5
remove edges (1,2) (3,4) (3,5)
remove edge (1,5)
2
1
3
4
5
2
1
3
4
5
remove edge (2,3)
2
1
3
4
5
SUBGRAPH A IS A TREE
SUBGRAPH B IS NOT A TREE

FIGURE 3. TEST FOR TREES


5
read data
fetch tree
convert to Fl. Pt.
tree non-zero
yes
calculate E. T. P. and power of s and accm. sum
1
no
last tree
yes D.P.
8
no
yes Trf.
6
1
D.P. or Trf.
Trf.
E.T.P. contains y_m
yes
set y_m to 1
store 2-tree
E.T.P. contains y_n
yes
2
D.P.
4
no
no
5
6
7
2
set y_n to 1
store 2-tree
last tree
yes
recover 2-tree and compare
like
check for 2tree sign
calculate E.2 T.P. and power of s and accm. sum
unlike
no
5
D.P.
5
PUNCH Y(S)
4
E.T.P. contains y_m
yes
7
no
5
8

E.T.P.—E.2 T.P. ELEMENTARY TREE (2-TREE) PRODUCT

FIGURE 4. Simplified Flow Diagram for Y(s) Driving Point and Transfer Admittance

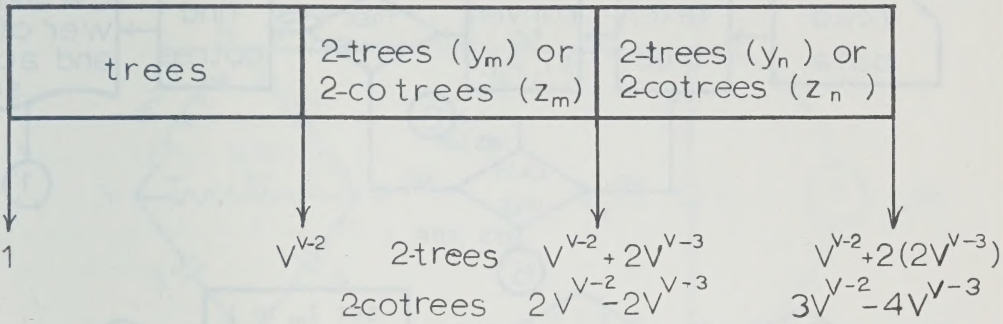

trees
2-trees (y_m) or
2-cotrees (z_m)
2-trees (y_n) or
2-cotrees (z_n)
1
V^{V-2}
2-trees
2-cotrees
$V^{V-2}+2V^{V-3}$
$2V^{V-2}-2V^{V-3}$
$V^{V-2}+2(2V^{V-3})$
$3V^{V-2}-4V^{V-3}$

FIGURE 5. Disk Storage

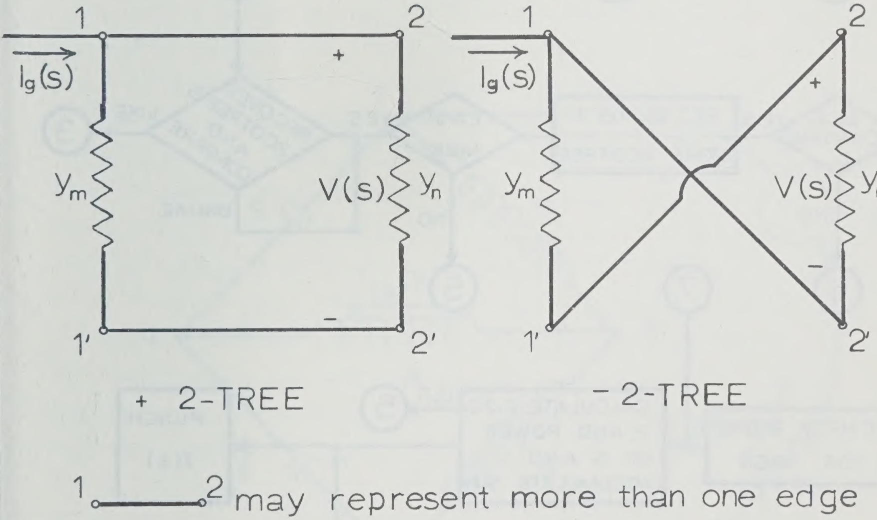

1
2
$I_g(s)$
+
y_m
V(s)
y_n
−
1'
2'
1
2
$I_g(s)$
+
y_m
V(s)
y_n
−
1'
2'
+ 2-TREE
− 2-TREE
1
2
may represent more than one edge

FIGURE 6. Representation for Determination of 2-tree Sign

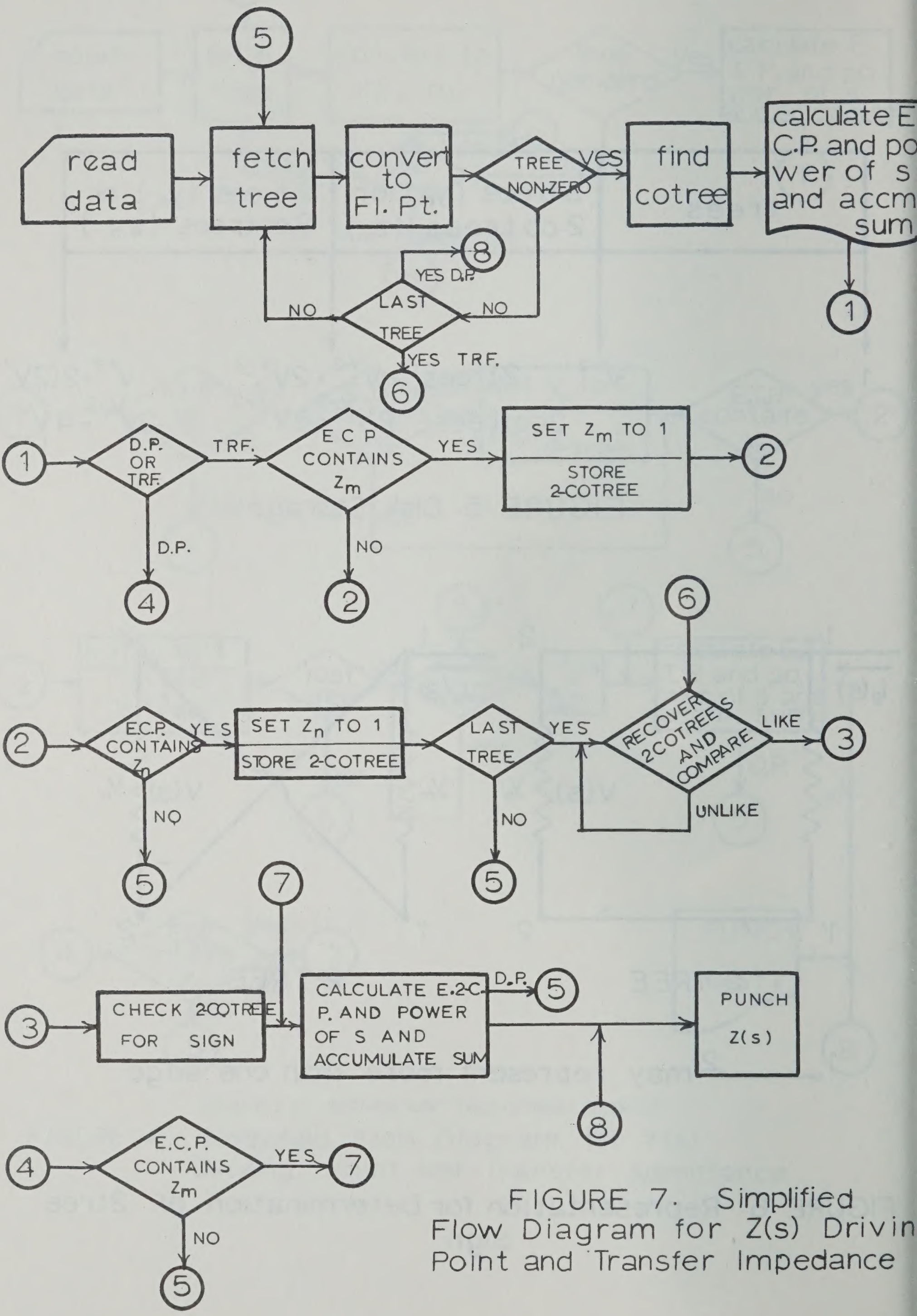
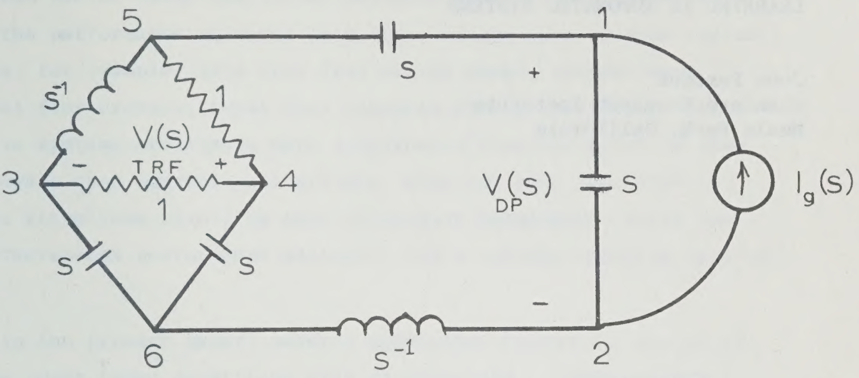
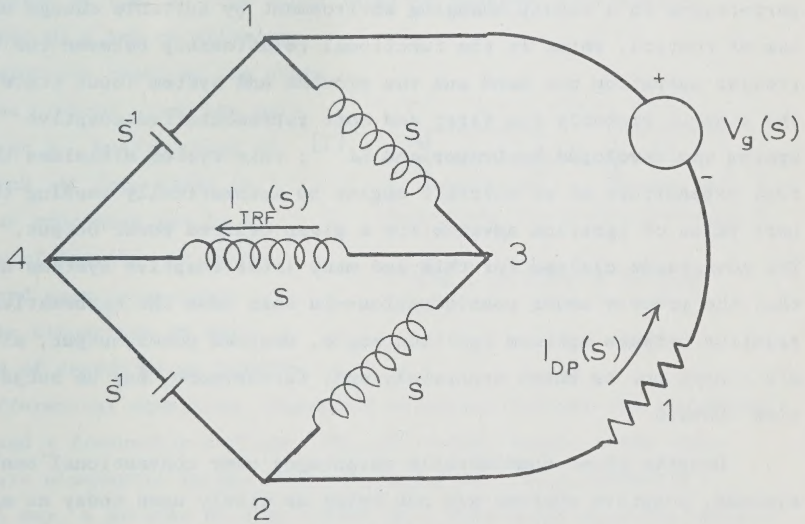

5
read data
fetch tree
convert to Fl. Pt.
TREE NON-ZERO
yes
find cotree
calculate E C.P. and po wer of s and accm sum
1
8
YES D.P.
LAST TREE
NO
NO
YES TRF.
6
1
D.P. OR TRF.
TRF.
E C P CONTAINS Z_m
YES
SET Z_m TO 1
STORE 2-COTREE
2
D.P.
4
NO
2
6
2
E.C.P. CONTAINS Z_n
YES
SET Z_n TO 1
STORE 2-COTREE
LAST TREE
YES
RECOVER 2COTREES AND COMPARE
LIKE
3
UNLIKE
NO
5
NO
5
7
3
CHECK 2-COTREE FOR SIGN
CALCULATE E.2-C. P. AND POWER OF S AND ACCUMULATE SUM
D.P.
5
PUNCH Z(s)
8
4
E.C.P. CONTAINS Z_m
YES
7
NO
5

FIGURE 7. Simplified Flow Diagram for Z(s) Drivin
Point and Transfer Impedance


5
1
S
+
S^-1
1
V(s)
-
TRF
+
3
4
1
S
S
V_DP(s)
S
I_g(s)
-
6
S^-1
2

(a) Network for Calculation of Y(s) Driving Point and Transfer Admittances


1
+
S^-1
S
V_g(s)
-
I_TRF(s)
4
3
S
I_DP(s)
S^-1
S
1
2

(b) Network for Calculation of Z(s) Driving Point and Transfer Impedances

FIGURE 8. Networks for Examples

LEARNING IN AUTOMATIC SYSTEMS

John Peschon
Stanford Research Institute
Menlo Park, California

A. INTRODUCTION

During the past decade, much research has been performed on adaptive systems; several definitions have been coined to differentiate adaptive systems from the more conventional control systems. The characteristic feature of the adaptive system is its ability to maximize performance in a slowly changing environment by suitable change of the law of control, which is the functional relationship between the controller output on one hand and the process and system input state on the other. Probably the first and most representative adaptive system was developed by Draper and Li[1]; this system minimizes the fuel expenditure of an aircraft engine by automatically seeking the best value of ignition advance for a given desired power output. The advantages claimed for this and many later adaptive systems are that the process under consideration--in this case the mathematical relation between optimum ignition angle, desired power output, altitude, etc.--need not be known accurately and, furthermore, may be subject to slow changes.

Despite these considerable advantages over conventional control systems, adaptive systems are not being as widely used today as may be expected. This may be due in part to the lack of analytical work and the number of diverse system concepts that work in a way but are probably far from being optimal designs.

In addition, adaptive systems have certain inherent limitations, primarily their sensitivity to measurement noise and their inability to track a rapidly changing environment, especially when more than two or three controller parameters have to be adjusted. Also, the performance measures used thus far are more or less instantaneous, for example, rate fuel flow or rms error, rather than integral (for example, total fuel consumed during the trip). Since adaptive systems are always more complicated from the point of view of hardware than conventional systems, they are best justified in complex situations involving many adjustment parameters, noisy and noninstantaneous performance measures, and a rapidly changing environment.

In the present paper, several approaches toward the design of systems under these conditions will be discussed. Since, in all cases, the control law presently used depends on past observations, we will refer to these systems as "learning."

B. FORMS OF LEARNING IN AUTOMATIC SYSTEMS

The automatic systems under consideration are represented by the diagram of Fig. 1. The block labelled *Process* is ordinarily described by a set of differential equations relating the state $\underline{x}$ to the control $\underline{u}$ and the perturbation $\underline{v}$. Descriptions of this kind not only apply to physical processes but to a variety of operational processes and even large-scale economic situations as well. Instead of descriptions involving differential equations, algebraic relations between the variables $\underline{x}$, $\underline{u}$, and $\underline{v}$ frequently suffice. In still other cases, these relations are stochastic rather than deterministic. In a perfectly general way, a process will be viewed as a functional relation between variables.

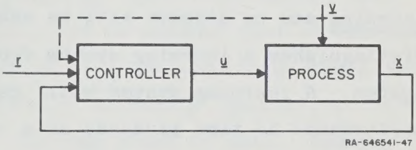


FIG. 1 AUTOMATIC SYSTEM

The block labelled "controller" involves another functional relation--referred to as the *law of control* or *control algorithm*--

between the command $\underline{r}$ and the state $\underline{x}$ and sometimes the measurable perburbations $\underline{v}$ and the control $\underline{u}$. In general this functional relation is algebraic rather than differential. The controller should be designed in such a way that the controlled system performs its function or objective in an optimal or near-optimal manner, despite changes in the environment as represented by $\underline{r}$ and $\underline{v}$.

The design of a satisfactory law of control often becomes an exceedingly difficult task, especially when the process under consideration is poorly understood, not stationary, and multivariable. Processes of this kind are frequently encountered in the operation of large-scale systems.

Instead of designing a law of control that is optimum or near optimum at the outset, there is a strong temptation to implement a law of control that, even though not optimal initially, has mechanisms built into it to ensure convergence toward optimal performance. A well-known example of this approach is the adaptive concept where the controller parameters are adjusted on the basis of a measurement of present performance.

There are various other ways whereby convergence of system performance toward the optimum can be implemented; these differ from the conventional adaptive approach. Such means will be referred to as learning and no attempt will be made here to define precisely what distinguishes a learning system from an adaptive or a conventional system. A learning system will, generally speaking, improve its performance as time proceeds as a result of a more efficient utilization of available past information than is usually the case.

In this paper three types of learning systems will be discussed and referred to as Type A, Type B, and Type C. These three types are believed to include all the systems referred to in the literature as "learning."

Type A, which is shown in Fig. 2, consists of observing an initially little-known process and of constructing a model thereof. The parameters of this model, which are measurable, then determine what the best settings of the controller should be. Type A evidently is only possible if the optimal relation between controller

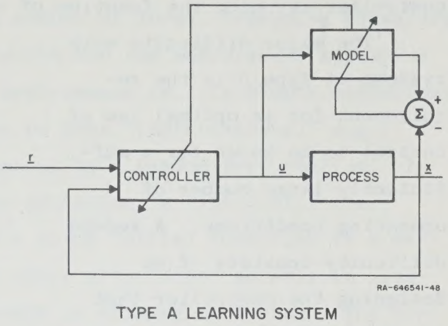


FIG. 2 TYPE A LEARNING SYSTEM

parameters and process parameters is known. The major difficulty lies in the construction of a sufficiently accurate model when the process may only be partially observable and when the observations may be corrupted by noise. There exist several techniques for model-making under those conditions, in particular the model parameters search scheme of Margolis and Leondes[2], the model parameter adjustment schemes discussed in Sec. D, and Kalman's[3] optimal filter. Even though Kalman's work is specifically concerned with the optimal--in the least squares sense--estimation of the state $\underline{x}$ when the process equations are exactly known; a series of recent papers extend this work to the case of process equations that are not accurately known[4,5].

The Type B system, shown in Fig. 3, contains a "trainer," which forces the controller to produce an output $\underline{u}$ equal to the trainer's output $\underline{u}^*$ for a finite number of operating conditions imposed upon the system during the training phase. It is assumed that an optimum --or at least satisfactory--control $\underline{u}^*$ is known for these conditions, and it is hoped that the controller will correctly "generalize" for all operating conditions once the training phase is over. Systems of Type B were first discussed by F. B. Smith[6], who implemented a time-optimal control by forcing $\underline{u}(t)$ to be equal to $\underline{u}^*(t)$ for a sufficient number of initial states $\underline{x}(0)$ in the error space. Similar systems have been described by Widrow[7]; a demonstration system called the "broom balancer" was developed by Widrow and co-workers. The controller is made up of "adaline" units*, which are adjusted during the training phase so as to mimic the action $\underline{u}^*$ of the trainer who attempts to hold a broom in its unstable upright position.

---

* Adaline is an adjustable network developed by Widrow[7].

Experiments were conducted with a human as well as with a predesigned controller assuming the function of the trainer.

The major difficulty with systems of Type B is the requirement for an optimal law of control to be known for a sufficiently large number of operating conditions. A second difficulty consists of so designing the controller that its parameters converge to those of the trainer, which are not always observable.

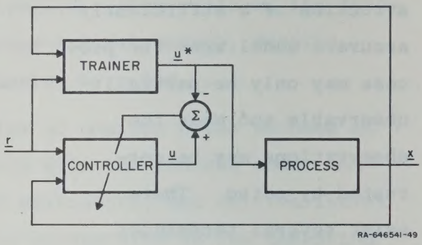


FIG. 3 TYPE B LEARNING SYSTEM SHOWN IN TRAINING PHASE
(The trainer is disconnected when training is over.)

The Type C learning system, shown in Fig. 4, is probably the most important from the point of view of practical application and also the most difficult to implement in meaningful situations. It does however come closest to an intuitive definition of learning, namely improvement of performance on the basis of past success or failure under similar circumstances.

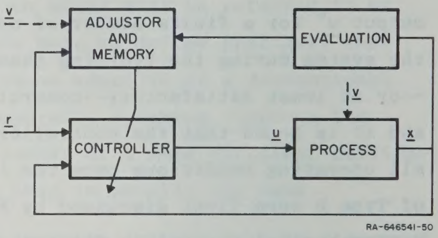


FIG. 4 TYPE C LEARNING SYSTEM
(The environmental perturbations, which may or may not define the system's operating condition, are shown by v.)

The Type C system comprises a performance evaluator connected to a mechanism for adjusting the controller parameters to improve performance. This mechanism in addition contains a memory wherein the past best adjustments for all operating conditions thus far encountered are stored. This last feature, which is not present in the systems ordinarily referred to as *adaptive*, leads to satisfactory performance in those situations where adaptive systems are not suited, i.e., where the performance evaluator and adjustment mechanism do not have sufficient time to produce optimum controller adjustment. This important feature will be discussed with greater detail in Sec. D-3.

The "evaluator" can take a number of forms, depending primarily on the performance criterion selected and the measurement apparatus available to determine how good performance is. In simple situations, the performance criterion is more or less "instantaneous," e.g., rate of fuel flow or rms error caused by a random perturbation. In more complicated situations, this criterion is "integral," i.e., reflects the cost of moving from a given initial condition to a desired final condition. In still other situations, notably in game-playing, the only result that counts is winning or losing. The continuous measurement of performance error--defined here as the difference between maximum performance and actual performance--may be possible, or else a "good-bad" statement regarding performance may be the only output available from the evaluator. In conventional adaptive systems, the performance criterion is almost always instantaneous and the measurement involves the change of performance δP caused by the change in a controller parameter $\delta\alpha$.

The design and performance of the adjustor and memory block largely depend on the nature of the performance criterion and on the output of the performance evaluator.

C. RELATED PAST WORK

Probably the first system (of Type C) discussed in the literature and demonstrated under the form of a maze runner is due to C. Shannon[8]. The maze is formed by partitioning a square board consisting of 25 squares. The runner is a sensing finger, which feels the partitions; this finger is moved in one of four directions by a set of two motors. The problem facing the maze runner is to get from some initial position to a final goal position, first by search and then by recall of the successful moves toward the goal. This constitutes the learning feature. In addition, various refinements are built into the machine, notably a forgetting feature required whenever the partitions or the goal is changed.

Despite its simplicity, Shannon's maze runner displays some very important features of learning systems, such as the possibility of limit-cycles--the runner following a "vicious circle" if allowed to by the partitions--or the use of a trajectory established in the past for a different initial condition.

In Ref [9], C. Shannon describes learning as follows:

"Suppose that an organism or a machine can be placed in, or connected to a class of environments and that there is a measure of "success" or "adaptation" to the environment. Suppose further that this measure is comparatively local in time, that is, that one can measure the success over periods of time short compared to the life of the organism. If this local measurement of success tends to improve with the passage of time, for the class of environments in question, we may say that the organism or machine is learning to adapt to these environments relative to the measure of success chosen..."

In Ref. [10] A. L. Samuel discusses machine learning in the game of checkers. Even though Samuel concerns himself with game-playing, his results are most relevant to learning in automatic systems where similar--and often simpler--algorithms can be used.

The checker program, which after various refinements succeeded in playing a championship game, contains several mechanisms for learning, only one of which will be discussed here. This mechanism centers around the "scoring polynomial," which attempts to measure the contribution to a final win of all possible sequences involving the twenty next moves. The correct coefficients of this scoring polynomial are not known initially, but are learned on the basis of success and failure. In order to get around the classical difficulty of selecting those coefficients of the scoring polynomial that should be increased and those that should be decreased, a most ingenious approach is taken. The computer--an IBM 704-- is programmed to include two players, α and β, which play against one another. The scoring polynomial for α, P_α, is slightly different from β's P_β. If α wins, P_α is presumably better than P_β. In the next game, β is given P_α and α is given a new polynomial P'_α, which, if better than P_α, is turned over to β in a third game. Conversely, if P_α is less effective than P_β, a search of the polynomial coefficients is made until α wins over β. Ultimately, the scoring polynomials of both α and β become optimal.

H. Zemanek[11] is probably the first author to discuss learning as applied to automatic systems in a general and descriptive manner. He distinguishes between

(1) The searching automaton,* which is essentially an adaptive system

(2) The searching automaton with memory, which is identical with the Type C system for instantaneous performance criterion

(3) The automaton with a fixed environment* model, where a law of control can be tried on a fast time scale (This approach is particularly well suited for integral performance criteria.)

(4) The automaton with learned environment model, which is identical to (3) except for an additional mechanism to improve the process model by observing how well it fits the real process

(5) The trained automaton, which is identical with the Type B system of the present paper.

In addition to the discussions by B. Widrow mentioned in Sec. B, there are several recent contributions dealing specifically with the subject of learning in conventional control systems. The work by Gibson, Fu, et al.[12] is primarily concerned with systems of Type C and instantaneous performance criteria. Several simple systems involving second-order servo-mechanisms have been simulated and shown to work. In addition, a system of Type C with an integral performance criterion was studied and shown to converge toward the optimal law of control. The system consists of a second-order relay control; the objective is to determine a switching line such that the weighted integral squared error

$$\int \underline{x}^T Q \underline{x} \, dt, \qquad (1)$$

evaluated between some initial state and the origin, is minimized. To do so, the $\underline{x}$ space is divided into measurement zones and probabilities p and 1-p of having relay outputs +1 and -1, respectively, are associated with each measurement zone. During experiments, these probabilities are reinforced or weakened depending on the least value of Eq. (1), evaluated over the past sample only. For measurement zones located outside the optimal switching line, p will converge toward 0 or 1; for measurement zones located on the switching line, p will converge toward an intermediate value. To overcome this difficulty, the initial measurement zone is further subdivided until p becomes either 1 or 0.

---

*Zemanek uses the terms "automaton" and "environment" for "automatic system" and "process."

E. C. Fraser, a co-worker of B. Widrow, similarly concerned himself with a relay system of unknown parameter values[13]. The object again is to find an optimal switching line. In a first experiment, the error state $\underline{x}$ (t) of a second-order system is displayed to a human judge on an X-Y plotter. The judge presses a "good" or "bad" button depending on whether he likes or dislikes the last segment of the $\underline{x}$ (t) trajectory. Fraser showed that the law of control converges toward an optimum if the judge is right in more than 50 percent of his decisions. He also related the speed of convergence to the judge's probability of being right.

In a second experiment, Fraser used a Type A system involving a model, the parameters of which are made to converge upon those of the process. These parameters then determine an optimal switching hyper-plane.

In addition to the above-listed references, several other papers should be mentioned, in particular Bush and Mosteller's stochastic learning models[14], which mathematically describe the learning performance of animals and which suggested the relay control of Gibson and Fu; D. W. Hagelberger's program is designed to play the game of matching pennies against a human opponent. The random process according to which the human selects heads or tails is modeled, with the result that the machine ultimately has a better than 50 percent change of winning[9]. The optimal filter of R. E. Kalman[3] and extensions thereof[4, 5] should also be included here, since optimal use is made of past information to estimate a state-variable and/or inaccurately known process parameter; when connected to a controller, this filter causes performance to improve with time.

None of the many papers concerned with neurological networks, that is special-purpose networks for adaptive pattern classification, were found to be relevant to the subject under discussion, since the system aspects are not touched upon in any depth. Much of the past work on adaptive systems, on the other hand, is directly applicable, since some of the system constituents (for example, the evaluator of Type C) are the same in both cases.

D. ADJUSTMENT ALGORITHMS

The major difficulty of any of the three learning schemes is to find convergent algorithms forcing the parameters of the model or of the controller toward their correct values. A general procedure to devise such algorithms does not exist. In the present section, several examples are treated and the convergence properties of the algorithms used are analyzed, mostly by means of Liapunov's second method.

In the first example, a model of a linear dynamic process, the state-variables and inputs of which are observable, is constructed.

In the second example, a tracking system of Type A is analyzed and its ultimate immunity to measurement noise is shown.

In the third example, a learning system of Type C is treated and its performance is compared to that of a conventional adaptive system.

Example 1: Modeling by Continuous Adjustment

The coefficients a^*_{ij} and $b^*_{k\ell}$ of the linear process

$$\dot{\underline{x}}^* = A^* \underline{x}^* + B^* \underline{u}, \tag{2}$$

are not known initially; the state $\underline{x}^*$ and the input $\underline{u}$ are measurable; it is desired to construct a model

$$\dot{\underline{x}} = A \underline{x} + B \underline{u} \tag{3}$$

and to adjust the coefficients a_{ij} and $b_{k\ell}$ such that they tend toward a^*_{ij} and $b_{k\ell}$. This is shown in Fig. 5.

To adjust the a_{ij} and $b_{k\ell}$, the following algorithm is suggested:

$$\begin{aligned} \dot{a}_{ij} &= -E_i \, x_j \\ \dot{b}_{ik} &= =E_i \, u_k \end{aligned} \tag{4}$$

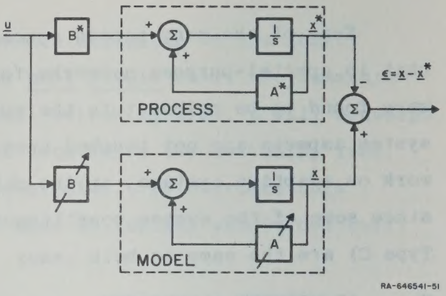


FIG. 5 MODELING OF A LINEAR PROCESS

To determine the conditions under which (4) constitutes a stable algorithm, the Liapunov function

$$V = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij} - a^*_{ij})^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{m}(b_{ik} - b^*_{ik})^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i - x^*_i)^2 \qquad (5)$$

is considered; using (4) $\dot V$ becomes

$$\dot V = -\sum_i\sum_j (a_{ij} - a^*_{ij})\epsilon_i x_j - \sum_i\sum_k (b_{ik} - b^*_{ik})\epsilon_i u_k + \sum_i \epsilon_i \dot\epsilon_i \ . \qquad (6)$$

If $\dot\epsilon_i$ is written as:

$$\begin{aligned}\dot\epsilon_i &= \dot x_i - \dot x^*_i = \sum_j a_{ij}x_j + \sum_k b_{ik}u_k - \sum_j a^*_{ij}x^*_j - \sum_k b^*_{ik}u_k \\ &= \sum_j a^*_{ij}x_j - \sum_j a^*_{ij}x_j + \sum_j a_{ij}x_j - \sum_j a^*_{ij}x^*_j + \sum_k (b_{ik} - b^*_{ik})u_k \\ &= \sum_j a^*_{ik}\epsilon_j + \sum_j (a_{ij} - a^*_{ij})x_j + \sum_k (b_{ik} - b^*_{ik})u_k, \qquad (7)\end{aligned}$$

the expression

$$\sum_i \epsilon_i\dot\epsilon_i = \sum\sum_{ij} a^*_{ij}\epsilon_i\epsilon_j + \sum\sum_{ij}(a_{ij} - a^*_{ij})\epsilon_i x_j + \sum\sum_{ik}(b_{ik} - b^*_{ik})u_k\epsilon_i \qquad (8)$$

can be substituted into Eq. (6) to provide

$$\dot V = \sum_i\sum_j a^*_{ij}\epsilon_i\epsilon_j \qquad (9)$$

The quadratic form (9) is negative-definite if the n symetric determinants Δ_ℓ, $\ell = 1, \ldots, n$

$$\Lambda_\ell = \begin{vmatrix} a^*_{11} & \ldots & a^*_{1\ell} \\ \vdots & & \vdots \\ a^*_{1\ell} & & a^*_{\ell\ell} \end{vmatrix} \qquad (10)$$

satisfy the inequalities:

$$\Delta_1 < 0;\ \Delta_2 > 0\ ;\ \ldots\ (-1)^n\,\Delta_n > 0$$

If the Liapunov function

$$W = \frac{1}{2}\sum_i x_i^{*2} \qquad (5a)$$

is used to investigate the stability of (2), $\dot W$ becomes

$$\dot W = \sum_i\sum_j a^*_{ij}x_i x_j \qquad (9a)$$

which is a quadric form identical with (9). Consequently, if (5a) is an admissible Liapunov function, the adjustment algorithm (4) is convergent.

Example 2: Tracking System

The object is to design a control system such that the angular position ψ (in either azimuth or elevation) of the tracking antenna coincides with the angular position of the target ψ^* despite the noise N entering into the monopulse difference channel and despite the temporal variations of ψ, which cause transient errors. The conventional approach toward tracking system design consists of minimizing--in the rms sense--the sum of both errors by inclusion of a suitable linear time-invariant filter into the loop. This approach largely discards all the known constraints placed upon the target's motion.

Frequently, the target's motion is governed by well-known laws, notably Newton's equations. It is hence reasonable to expect that tracking errors could be progressively reduced, as more knowledge is gained about ψ^* as a result of past observation; that is, as a result of modeling the target trajectory. A scheme to do so is shown in Fig. 6.

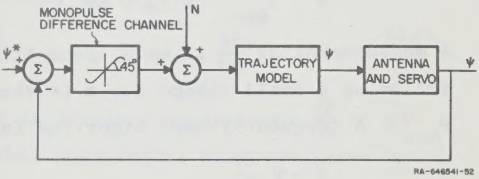


FIG. 6 TRACKING BY MODELING THE TARGET TRAJECTORY
(ψ^* = target angle, ψ = antenna angle)

The true motion ψ^* is assumed to be

$$\psi^* = \psi^*(o) + v^* t \quad (11)$$

and the motion ψ given by the adjustable model is

$$\psi = \psi(o) + v t \quad (12)$$

If the antenna and servo dynamics can be neglected and if the tracking errors remain within the linear region of the monopulse difference channel, the model adjustment algorithm

$$\dot{v} = g(\psi^* - \psi + N) \qquad g = \text{constant} \qquad (13)$$

leads to the time-varying differential equation

$$\dot{v} + g v t = g[\psi^*(o) - \psi(o)] + g v^* t + g N \quad . \qquad (14)$$

Equation (14) can be solved in terms of the normal integral, which is tabulated; the following results can be proved readily for $g > o$:

(1) As $t \to \infty$, $v \to v^*$ and the tracking error $\psi - \psi^*$ tends toward zero.

(2) As $t \to \infty$, the noise-induced error tends to zero; for N gaussian, white, of zero mean and power σ_N^2, the mean squared tracking error is approximately $(g\sigma_N^2)/(2t)$.

(3) If the target trajectory is not exactly as in (11), for example, if there is an acceleration term $(1/2)(\gamma^* t^2)$, the resulting tracking error is bounded.

Example 3: A Type C system with instantaneous performance criterion

Since the Type C system in its simplest form appears similar to a conventional adaptive system, both systems are analyzed and the superiority of the Type C system is demonstrated for instantaeous performance criterion.

a. Adaptive System

Let the performance P of a process depend on an adjustable parameter α and a measurable environmental parameter β. The relation between P, α, and β is not known beforehand.

In the vicinity of the optimum point $[P_{max}, \hat{\alpha}]$, P varies as

$$P = P_{max} - A(\alpha - \hat{\alpha})^2 \qquad A > 0 \tag{15}$$

A measurement of $\frac{\partial P}{\partial \alpha}$ is made, generally by observing the change δP following a small change $\delta\alpha$; α is then adjusted to drive P toward P_{max}. A frequently used algorithm is:

$$\dot{\alpha} = K \frac{\partial P}{\partial \alpha} \tag{16}$$

Since

$$\frac{\partial P}{\partial \alpha} = -2A(\alpha - \hat{\alpha}) \quad , \tag{17}$$

the adaptive system is described by the linear differential equation

$$\dot{\alpha} = -2KA(\alpha - \hat{\alpha}) \quad , \tag{18}$$

the steady-state solution of which is

$$\alpha = \hat{\alpha} \tag{19}$$

for constant $\hat{\alpha}$ and provided that $2KA > 0$.

The measurement $\frac{\partial P}{\partial \alpha}$ is always affected by noise N, which enters into Eq. (18) as

$$\dot{\alpha} = -2KA(\alpha - \hat{\alpha}) + KN \tag{20}$$

If N is white, gaussian, of zero mean and power σ_N^2, the resulting mean square misadjustment σ_α^2 is readily calculated from (20)

$$\sigma_\alpha^2 = \frac{K}{4A}\sigma_N^2 \tag{21}$$

and the average loss in performance due to noise is

$$P_{max} - P = \frac{K}{4}\sigma_N^2 \tag{22}$$

It is clear from Eq. (22) that the gain K is limited by the average performance error $\overline{P_{max} - P}$ deemed admissible. Since $\hat{\alpha}$ is assumed to be a function of the environmental parameter $\beta(t)$, i.e., $\hat{\alpha} = \hat{\alpha}(t)$, there will be a tracking error in addition to the noise induced error. For a known variation $\hat{\alpha}(t)$, this tracking error is determined readily from Eq. (20). If the power spectrum of $\hat{\alpha}(t)$ is known, the gain K can be selected to minimize the combined effect of tracking and noise-induced errors in accordance with the well-known techniques of optimum filter design. The important point is that, unless $\hat{\alpha}$ is a constant, there will always be tracking errors and the performance error $P_{max} - P$ will depend on both N and β.

b. Type C System

Let it be assumed that $\hat{\alpha}(t)$ is linearly related to the measurable environmental parameter $\beta(t)$ as in

$$\hat{\alpha}(t) = \hat{C}\,\beta(t) \tag{23}$$

where $\hat{C}$ is initially unknown.

Instead of utilizing the performance measurement $\frac{\partial P}{\partial \alpha}$ to adjust α, as was done in Eq. (16), this measurement will now serve to adjust the variable gain C, as shown in Fig. 7. With the adjustment algorithm,

$$\dot{C} = K'\frac{\partial P}{\partial \alpha} \tag{24}$$

and the misadjustment $C - \hat{C}$ defined as ΔC, the system will then be described by the time-varying differential equation

$$\Delta\dot{C} = -2K'A\beta(t)\Delta C + K'N \quad , \tag{25}$$

which, for N = 0, integrates to

$$\Delta C = \Delta C(o)e^{-2K'A\int_o^t \beta(t)dt} \quad . \tag{26}$$

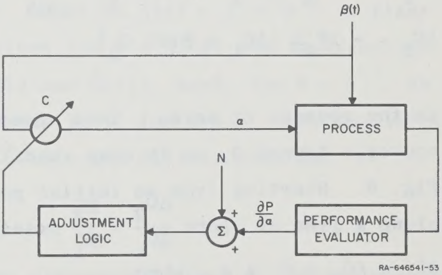


FIG. 7 TYPE C SYSTEM

The misadjustment ΔC converges toward zero if $K'A\int_o^\infty \beta(t)dt$ tends to ∞ as $t \to \infty$. The performance error

$$P_{max} - P = A[\Delta C(o)]^2 \beta^2 e^{-2K'A\int_o^t \beta(t)dt} \qquad (27)$$

caused by tracking errors hence converges toward zero as the initially unknown relation between $\hat{\alpha}$ and β is learned by the system.

Since for slowly varying $\beta(t)$, the mean squared noise induced misadjustment σ_C^2 is

$$\sigma_C^2 = \frac{K'}{4A\beta(t)} \sigma_N^2 \quad . \qquad (28)$$

A somewhat obvious refinement of this type C system is to reduce K' progressively with the result that both tracking and noise-induced errors will tend to zero as $t \to \infty$.

The assumed relation between $\hat{\alpha}$ and $\beta(t)$--see Eq. (23)--is perhaps not a realistic one, but it displays the properties of the resulting system. A better relation, valid for any continuous function $\hat{\alpha} = f(\beta)$ in the vicinity of the operating point β_o is

$$\hat{\alpha} = f(\beta_o) + \frac{\partial f(\beta_o)}{\partial \beta}(\beta - \beta_o) = \hat{C}_1 + \hat{C}_2 \beta \quad , \qquad (29)$$

where $\hat{C}_1$ and $\hat{C}_2$ are constants.

The adjustment algorithm

$$\dot{C}_1 = K' \frac{\partial P}{\partial \alpha} \qquad \dot{C}_2 = K'_2 \frac{\partial P}{\partial \alpha} \qquad (30)$$

leads to the closed loop system equations

$$\begin{aligned} \Delta\dot{C}_1 &= -\,2K'_1A\,[\Delta C_1 + \beta(t)\Delta C_2] \\ \Delta\dot{C}_2 &= -\,2K'_2A\,[\Delta C_1 + \beta(t)\;C_2] \end{aligned} \qquad (31)$$

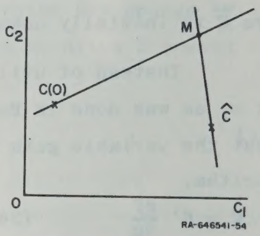


FIG. 8 PHASE PLANE DIAGRAM OF EQ. (31)

in the absence of noise. This algorithm does not cause ΔC_1 and ΔC_2 to converge toward 0, as is seen readily from the phase plane diagram of Fig. 8. Starting from an initial point C(0), motion will proceed along a line of slope $\frac{\Delta C_2}{\Delta C_1} = \frac{K'_2}{K'_1}$ passing through C(0) and will stop when $\Delta\dot{C}_1 = \Delta\dot{C}_2 = 0$; i.e., when

$$\Delta C_1 + \beta(t)\; \Delta C_2 = 0 \quad , \qquad (32)$$

which is satisfied along the straight line $\hat{C}M$ of slope $-\frac{1}{\beta(t)}$. The equilibrium, which is the intersection of the two lines at M, shifts back and forth along C(0)M as β varies; it will not end up at $\hat{C}$, unless $\hat{C}$ is located on C(0)M, which is a special case.

A general procedure for selecting a convergent algorithm in situations where a single error signal is used to adjust several independent variables was found by E. C. Fraser [13]. Let

$$\begin{aligned} \Delta\dot{C}_1 &= K' \frac{\partial P}{\partial \alpha} \\ \Delta\dot{C}_2 &= K'\beta(t) \frac{\partial P}{\partial \alpha} \end{aligned} \tag{33}$$

and consider the closed-loop system equations

$$\begin{aligned} \Delta\dot{C}_1 &= -2K'A[\Delta C_1 + \beta(t)\Delta C_2] \\ \Delta\dot{C}_2 &= -2K'A\beta(t)[\Delta C_1 + \beta(t)\Delta C_2] \end{aligned} \tag{34}$$

The convergence properties of Eq. (34) are readily analyzed by means of Liapunov's second method. Let

$$V = \tfrac{1}{2}(\Delta C_1^{\,2} + \Delta C_2^{\,2}) \tag{35}$$

$$\dot{V} = -2K'A(\Delta C_1 + \beta\Delta C_2)^2 \quad . \tag{36}$$

$\dot{V}$ is clearly negative definite for $2K' > 0$ unless

$$\Delta C_1 = \Delta C_2 = 0 \tag{37}$$

or $$\Delta C_1 = -\beta(t)\Delta C_2 \tag{38}$$

Case (37) is admissible; Case (38) signifies that the motion will settle at point M if $\beta(t)$ is constant for a sufficiently long period of time. However, as soon as a change in β occurs, the quantity $(\Delta C_1 + \beta\Delta C_2)^2$ becomes positive, $\dot{V}$ becomes negative and the distance from the equilibrium point $\hat{C}$ must diminish. Hence, unless β = constant, C will converge toward $\hat{C}$ at a rate which depends, among others, on the variations $\beta(t)$. The mechanism whereby convergence occurs is illustrated in the phase-plane diagram of Fig. 9.

Assuming a piecewise constant function $\beta(t) = \beta'$ or β'', a first equilibrium condition C(1) is obtained for $\beta = \beta'$ at the intersection of the perpendicular lines C(0) C(1) and $\hat{C}$C(1); next, for $\beta = \beta''$, an equilibrium condition C(2) is reached; last, for $\beta = \beta'$, the equilibrium condition C(3) which is closer to $\hat{C}$ then was C(1) is obtained.

E. IMPLEMENTATION

For all three types of systems discussed in Sec. B, an adjustable structure -- process model or controller -- is required. There exist three general approaches toward implementing such structures, namely:

(1) The general-purpose digital computer,

(2) The so-called neurological nets, and

(3) Special-purpose analog structures.

Because of the extraordinary versatility of the general-purpose digital computer, any adjustable structure can be implemented on it, in principle, although its capacity may be exceeded or its cost not justified for certain applications.

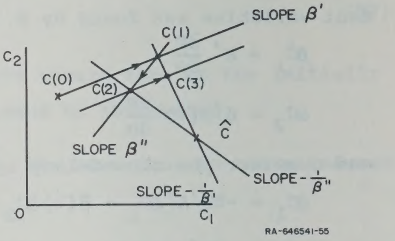


FIG. 9 PHASE PLANE DIAGRAM OF EQ. (34)

Neurological nets have been widely discussed in the recent past in connection with pattern recognition and pattern classification. There exist several similar--from the system point of view--implementations, in particular Rosenblatt's Perceptron [15], Widrow's Adaline [7], Steinbuch's Lernmatriz [16], Rosen and Nilsson's Minos [17], Smith's adjustable digital net [6], and others. Most of the neurological nets are made of "threshold logic units" (TLU), which consist of adjustable weights w_i, a summary junction and a relay-type nonlinearity, as shown in Fig. 10.

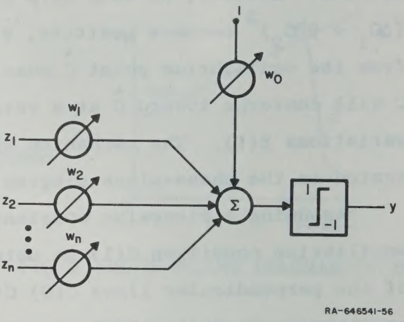


FIG. 10 THRESHOLD LOGIC UNIT (TLU)

The input pattern $\underline{z}$ consists of the binary (either 1 or 0) numbers z_i; depending on the weights w_i, the output is either +1 or -1. The attractive feature of the TLU is the systematic way in which adjustment of the weights takes place. Suppose that for a given pattern $\underline{z}$, y is -1, but should be +1. Then, all the weights connected to a +1 input are incremented by the same amount Δw and all those connected to a 0 input are decremented by Δw until y becomes +1. This procedure is repeated for all the patterns $\underline{z}$, $\underline{z}'$, $\underline{z}''$. . . which the TLU is required to separate into +1 and -1 categories, It is known that if there exists a weight vector w^* that correctly separates the patterns $\underline{z}$, $\underline{z}'$, $\underline{z}''$ . . ., then the adjustment procedure will force any arbitrary initial weight vector $\underline{w}(o)$ toward $\underline{w}^*$. The conditions under which $\underline{w}^*$ exist are also known (linear separability) [17].

More complicated neural net structures are made up of a number of TLU's, as shown in Fig. 11.

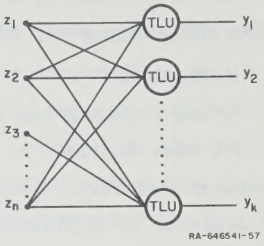


FIG. 11 SINGLE-LAYER INTERCONNECTION OF TLU's

The network of Fig. 11 can, in principle, be used to implement various algebraic relations between an analog input $\underline{u}$ and an analog output $\underline{x}$. Each input variable $\underline{u}_l$ is assigned a code $\underline{z}_\ell$ consisting of $z_{\ell 1}, z_{\ell 2}, \ldots z_{\ell n}$ and each output variable x_j is the result of decoding the pattern $\underline{y}_j$ consisting of $y_{j1}, y_{j2}, \ldots y_{jk}$. If the codes are such that a correct weight vector $\underline{w}^*$ exists, then an initially incorrect vector $\underline{w}(o)$ will converge toward $\underline{w}^*$, when the above described adjustment procedure is applied. This procedure does require that the error vector $\underline{\varepsilon}_{j_*} = \underline{y}_j^* - \underline{y}_j$, when $\underline{y}_j^*$ is the output pattern associated with the output x_j, be available.

The code associated with u_ℓ and x_j is not arbitrary, but must be chosen in such a manner that the relation between u_ℓ and x_j can in fact be implemented. Permissible input codes have been studied by F.W. Smith [6].

If an input variable u_ℓ is quantized into seven levels (0.78 percent accuracy) and x_j is quantized into six levels (1.56 percent accuracy), the number of weights required is $7 \cdot 6 = 42$. Even though this is not a large number, it is still a consideration in view of the relatively cumbersome and costly weighting elements (magnetic cores and chemical elements) now in use.

These neurological networks are claimed to possess high reliability even though the weights may not be very reliable, because the function of a faulty weight is at least partially taken over by its neighbors. They also possess the ability to "generalize", that is, extrapolate and interpolate for input patterns not seen during the training phase.

As a third possibility to provide adjustable structures suitable for automatic control, a variety of linear and nonlinear analog devices are available, but their use has not been considered in a general way. Suppose that the functional relation between x_j and u_ℓ is of the form

$$x_j = a_o + a_1 u_\ell + a_2 u_\ell^2 + \ldots \qquad (39)$$

where the a_1 are initially unknown. A relation such as Eq. (39) is readily implemented with reasonable accuracy (up to 1 percent) by the analog structure shown in Fig. 12.

F. APPLICATION

Since learning has as yet not been included in any practical automatic system, a discussion of applications is of necessity speculative, but perhaps not futile in view of the excellent learning programs in existence for game-playing.

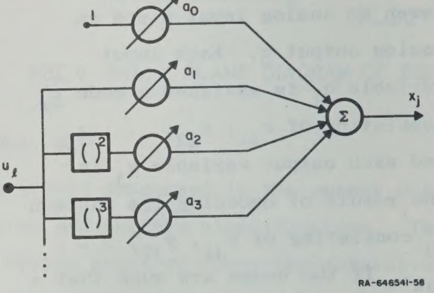


FIG. 12 ADJUSTABLE ANALOG STRUCTURE

The major limitation of the Type A approach is the requirement that the optimum relation between process parameters and controller parameters must be known beforehand. At first sight, this limitation appears to restrict the use of the Type A approach to relatively simple processes the parameters of which are subject to large and not directly measurable variations with time. In fact, there exist many special processes, where this approach is promising, as evidenced by the tracking system described in Sec. D-2.

Another application of the Type A approach is Hagelberger's modeling of the statistics governing one player's moves, as mentioned in Sec. C for a head or tail game. This technique can clearly be extended to the practically more important case of game theory.

The major shortcoming of the Type B approach is the need for a trainer, that is, for good law of control to be known in a sufficiently large number of conditions. There appear to be two classes of control problems where this is not a major shortcoming.

The first class comprises all those situations where a skilled human is available for training and where, for some reason, it is desirable to replace this human by a control mechanism. Situations of this kind range from the imitation of a skilled machinist's motion for reasons of economy to the automatic execution of emergency procedures, trained into the system during realistic drills, to avoid inappropriate decisions as a result of panic or pressure.

The second class involves situations where the solution of the control problem requires an extensive amount of computation, as in most optimal control problems. Training in this case is performed by a suitable computer, possibly in slow time. After the controller has been adjusted, the computer can be disconnected and turned over to other tasks. This is precisely the approach described in [6].

As for Type C systems, applications are to be found whenever the conditions, for which a particular law of control is best, are recurrent, or when the computations required to evolve an optimal law are lengthy, as is generally the case with integral criteria.

As an example of a very complex recurrent situation, the dynamic process represented by a city with its pattern of streets, intersections and cars may be considered. At any one time, the state of this process is described by the number of cars contained in each lane between intersections. Actuation of the traffic lights changes the state in a desirable or undesirable way depending on how well the red-green sequence of each traffic light is related to the state of traffic, not only of its intersection, but also of that of adjacent intersections. The optimum[*] sequence furthermore depends on the environment: weather, time of day, etc. At present, it is not possible to program a control computer to generate an optimum sequence for more than a few lights. One might attempt to incorporate conventional adaptation into the program by altering slightly the sequence of successive lights and by observing the resulting change of performance upon the entire system. Since traffic changes radically within fractions of an hour, this conventional approach is clearly not appropriate.

In order to overcome this difficulty the major environmental parameters -- time of day, weather etc. -- are stored together with the last best settings found for each traffic light exactly as in Example 3 of Sec. D. The next time this same environment prevails, the search will use as initial conditions these last best settings retrieved from the memory. Ultimately, the memory will contain the optimum settings associated with each environmental condition and the search can be discontinued.

---

[*] A possible definition of a relatively instantaneous optimum might be the intersection throughput averaged over all intersections.

A similar system of Type C with instantaneous performance criterion, search and storage of environmental parameters for various steel mill applications, is proposed in [18].

As for Type C systems with integral criterion, the optimal control problem solved via dynamic programming is considered. If the number of state-variables exceeds three or four, conventional dynamic programming overtaxes present-day computers, when an accurate optimum trajectory is desired at the first attempt. As an alternative, the $\underline{x}$, $\underline{u}$, and t space are first quantized coarsely and an approximate optimal trajectory is obtained and stored. The next time an optimal trajectory is desired for adjacent initial (and possibly) final states, quantization levels are reduced and the search is limited to a narrow tube about the previous optimal trajectory. This results in a better optimum, which in turn is stored. Proceeding in this manner, and using state-increment dynamic programming [19], processes of high order can be optimized without overtaxing the computer.

G. ACKNOWLEDGMENT

The assistance of Mr. E. C. Fraser of Stanford University and Stanford Research Institute and Dr. R. E. Larson of Stanford Research Institute in the preparation of the present paper is gratefully acknowledged.

H. REFERENCES

1. Draper, C. S. and Li, Y. T., "Principles of Optimalizing Control Systems and Applications to the Internal Combustion Engine," ASME publication, September 1951.

2. Margolis, M. and Leondes, C. T., "A Parameter Tracking Servo for Adaptive Control Systems," University of California, Department of Engineering, August 1959, p. 30.

3. Kalman, R. E. and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," Trans. ASME, 83 D, pp.95-108, (March 1961).

4. Schmidt, S. F., "State Space Techniques Applied to the Design of a Space Navigation System," JACC, (1962).

5. Mayne, D. Q. "Optimal Nonstationary Estimation of the Parameters of a Linear System with Gaussian Inputs," Journal of Electronics and Control, Vol. 14, No. 1, pp. 101-112, (January 1963).

6. Smith, F. B., "A Logical Net Mechanization for Time-Optimal Regulation," NASA Technical Note D-1678, p. 34, (December 1962).

7. Widrow, B., "Generalization and Information Storage in Networks of Adaline Neurons," Self-Organizing Systems, Yovits (Ed.), pp. 435-463, (Spartan Books, Washington, D.C., 1962).

8. Shannon, C. E., "Presentation of a Maze Solving Machine," 8th Conference on Cybernetics, pp. 173-180, (1951).

9. Shannon, C. E., "Computers and Automata," Proc. IRE, pp. 1234-1241, (October 1953).

10. Samuel, A. L. "Some Studies in Machine Learning Using the Game of Checkers," pp. 71-108, Feigenbaum, E. A. and Feldman, J. (Ed.) Computers and Thought, (McGraw-Hill, 1963).

11. Zemanek, H., "Lernende Automaten," Taschenbuch der Nachrichtenverarbeitung, (Springer 1962).

12. Gibson, J. E., et. al., "Philosophy and State of the Art of Learning Control Systems," Purdue University, School of Electrical Engineering, (November 1963).

13. Fraser, E. C., "Self-Stabilizing Adaptive Control Systems," Technical Report in preparation, Stanford Electronics Laboratories, Stanford, California, (1964).

14. Bush, R. R. and Mosteller, R., Stochastic Models for Learning, p. 365, (John Wiley, 1955).

15. Rosenblatt, F., Principles of Neurodynamics - Perceptrons and the Theory of Brain Mechanisms, (Spartan, Washington, D.C., 1961).

16. Steinbuch, K., "Schaltungen mit der Lernmatriz," Lernende Automaten, p.63-68, (R. Oldenbourg, München, 1961).

17. Nilsson, N. J., Learning Machines - Foundations of Trainable Pattern Classifying Systems, (McGraw-Hill, 1965).

18. Kussl, V., "Lernende Automaten in Hüttenwerken," p. 187-205, Lernende Automaten, (R. Oldenbourg, München, 1961).

19. Larson, R. E., "State Increment Dynamic Programming," Allerton Conference, University of Illinois, 1964.

ON THE LEARNING BEHAVIOR OF FINITE-STATE SYSTEMS IN RANDOM ENVIRONMENTS*

G. J. McMurtry and K. S. Fu

School of Electrical Engineering
Purdue University
Lafayette, Indiana

I. INTRODUCTION

The automata concept is sufficiently broad to be applicable to many different types of systems, serving as a basis for either analysis or synthesis. The theory of finite automata has been applied rather extensively to such areas of study as circuit and information theory and computer design. This paper is concerned with an application of automata theory to a class of learning control systems.

The behavior of finite-state deterministic and probabilistic automata operating in a random environment has been investigated by Tsetlin(1) and Varshavskii and Vorontsova.(2) They also considered the case of a random environment whose statistical properties obey a Markov chain. A learning control system(3) is considered to be one which not only adapts itself to each new environment, but is also able to recognize the environment, recall from experience the best

---

* This work was supported in part by the National Science Foundation Grant GP-2183.

compensation, implement it, and continue to adapt. An informal introduction to learning control systems has been given by Fu(4) and a recent paper(5) reports on a class of learning systems based on the above philosophy.

In this paper the automaton represents that part of the learning controller which performs the adaptive logic. The basic model of Tsetlin, which is found to require certain restrictions, is studied as an optimum-seeking finite-state automaton operating in an adaptive manner in a random environment. A modified automaton is proposed which changes its structure as it adapts on each environment, and the controller stores in memory the last automaton structure for that environment.

II. BASIC MODEL AND SYSTEM

A. BASIC MODEL

Consider a deterministic automaton, A, operating in a random environment, C, as shown in Figure 1. The state, input, and output sets of A are $Q = (s_1, \ldots, s_t)$, $X = (x_1, x_2)$, and $Y = (y_1, \ldots, y_K)$, respectively. The operation of A is deterministic and described by

$$s(t+1) = F[s(t), x(t)]; \; y(t) = G[s(t)] \qquad (1)$$

For each input, x, the state transitions may be described by a transition matrix, A(x), with elements

$$a_{ij}(x) = P[s(t+1) = s_j/s(t) = s_i \;, \; x(t) = x] \qquad (2)$$

Since A is deterministic, each row in A(x) contains one, and only one, element equal to unity for a given x. All other elements are zero.

The random environment, $C = C(p_1, \ldots, p_K)$, is stationary and responds to each output, y_i, as follows:

$$p_i = P[x(t) = 1/y(t) = y_i] \quad (i = 1, \ldots, K) \tag{3}$$

$$q_i = 1-p_i = P[x(t) = 0/y(t) = y_i] \quad (i = 1, \ldots, K) \tag{4}$$

where $x_1 = 0$ and $x_2 = 1$ represent a non-penalty or penalty, respectively. Then

$$p_{ij} = P[s(t+1) = s_j/s(t) = s_i] = p_\alpha a_{ij}(1) + q_\alpha a_{ij}(0) \quad (i = 1, \ldots, t) \tag{5}$$

where $y_\alpha = G[s_i]$. Since only one element, $a_{ij}(x)$, equals unity in row j of A(x)

$$\sum_{j=1}^{t} p_{ij} = p_\alpha + q_\alpha = 1 \quad (i = 1, \ldots, t) \tag{6}$$

and the matrix P, with elements p_{ij}, is stochastic (all rows sum to unity). Thus the automaton operating in a stationary random environment is represented by a Markov chain.

The transition diagram for the automaton with "linear tactic", as defined by Tsetlin, with k branches and n states in each branch, is shown in Figure 2. The state set of A is

$$Q = (s_{11}, s_{12}, \ldots, s_{1n}, s_{21}, s_{22}, \ldots, s_{k-1,\,n}, s_{k1}, \ldots, s_{kn})$$

where the subscripts represent the branch in which the state lies and the position in the branch, respectively. If $s(t) = s_{ij}$, the state transitions are given by

$$s(t+1) = \begin{cases} s_{i,\,j+1}/j \neq n;\ s_{in}/j = n : x(t) = 0 \\ \\ (s_{i,j-1}/j \neq 1),\ (s_{i+1,\,1}/i \neq k;\ s_{11}/i = k : j=1) : x(t)=1 \end{cases} \tag{7}$$

The output set is $Y = (y_1, y_2, \ldots, y_k)$ where

$$y(t) = y_i = G[s(t) = s_{ij}] \quad 1 \leq i \leq k,\ 1 \leq j \leq n \tag{8}$$

A performance criterion for the automaton is defined as the expected value, M, of the penalty. If the Markov chain is ergodic then for the automaton with linear tactic

$$\lim_{n \to \infty} M = M_{min} = \min\,(p_1, \ldots, p_k). \tag{9}$$

This assumes, in addition to ergodicity, that at least one $p_i < 0.5$ so that the automaton tends to become imbedded in the branch with $p_i = M_{min.}$ In addition, it is shown below that one, and only one, p_i must be less than 0.5 for convergence to M_{min}.

Varshavskii and Vorontsova considered the same problem except that A is probabilistic. The transition probabilities between states are adjusted and normalized by reinforcement techniques, both linear and quadratic, using an arbitrary adjustment parameter.(6)

Krylov(7) modified Tsetlin's model by making the automaton pseudo-probabilistic. Transition probabilities when $x(t) = 1$ are

$$\begin{aligned} P[s(t+1) = s_{i,j-1}/s(t) = s_{ij}\ ;\ j \neq 1] &= P[s(t+1) = s_{i,j+1}/s_{ij}\ ;j \neq n] \\ &= P[s(t+1) = s_{i+1,\,1}/s(t) = s_{i1}\ ;\ i \neq k] \\ &= P[s(t+1) = s_{in}/s(t) = s_{in}] = P[s(t+1)=s_{11}/s(t) = s_{k1}]=0.5 \end{aligned} \tag{10}$$

If $x(t) = 0$, the transitions occur as in the linear tactic model. This gives the effect of halving each p_i and optimum behavior results, even with all $p_i > 0.5$.

B. MARKOV ASPECTS

Assume now that in the linear tactic model k = 2. Then the Markov transition matrix is:

$$P = \left[\begin{array}{cccccccccccc} s_{1n} & s_{1,n-1} & s_{1,n-2} & \cdots & s_{12} & s_{11} & s_{21} & s_{22} & \cdots & s_{2,n-2} & s_{2,n-1} & s_{2n} \\ q_1 & p_1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ q_1 & 0 & p_1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & q_1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & q_1 & 0 & p_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & p_2 & 0 & q_2 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & p_2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & q_2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & p_2 & 0 & q_2 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & p_2 & q_2 \end{array}\right] \tag{11}$$

Thus P is a random walk with reflecting barriers$^{(8)}$ at states s_{1n} and s_{2n}.

Assume that $p_2 < p_1 < 0.5$ and the initial state is s_{11}. It seems obvious that there is a finite probability, σ, that the automaton never reaches s_{21}, and therefore never enters the branch which yields $p_2 = M_{min}$. It can be shown$^{(9)}$ that as $n \to \infty$, $\phi = \frac{p_1}{q_1}$ represents the probability of ever entering state s_{21}. Thus

$$\sigma = 1 - \phi = \frac{q_1 - p_1}{q_1} > 0 \tag{12}$$

is the probability that state s_{21} is never reached and $\lim_{n \to \infty} M \neq M_{min}$. This condition is demonstrated in the results discussed below.

C. APPLICATION TO A CLASS OF LEARNING CONTROL SYSTEMS

The basic system used for application of the above automata is shown in block form in Figure 3. The input, R, is assumed to consist of steps applied at discrete intervals, sufficiently spaced so that a good evaluation of the I.P. may be made during each step. In the system discussed here, I.P. $= \int_0^\infty e^2 dt$. The I.P. is compared to a threshold, θ, and

$$x(t) = \begin{cases} 1 & \text{I.P.} > \theta \\ 0 & \text{I.P.} \leq \theta \end{cases} \qquad (13)$$

The input to A may be considered as an operator which reinforces the actions of A that yield acceptable performance.

The parameter b of the controlled plant is assumed to be measurable and remain constant during each control interval. If b varies, learning is introduced by remembering optimum settings for each measurable value of b. The parameter, $\underline{a}$, is set by the controller through a transducer, AA, and is subject to unknown disturbances, d. The admissible values of $\underline{a}$ and y are quantized into discrete levels and the controller learns the optimum y. In the system simulated on the IBM 7090, $y = \underline{a}$.

Now if the automaton, A , is a learning controller and C represents a controlled plant subject to random disturbances, then Figure 1 is a simplified description of the basic learning control system. In obtaining the results below, the linear tactic model was used as the learning controller, with $k = 6$ and $2 \leq n \leq 20$. The value of b was fixed at $b = 9$. It may be shown(9) that the optimum undisturbed setting of $\underline{a}$ is $\underline{a} = \sqrt{b}$, which for this problem yields $\underline{a} = 3$ and $\text{I.P.}_{min} = 0.333$. The output set was chosen as $Y = (1, 2, \ldots, 6)$, i.e., $y_i = i$, and $0.334 \leq \theta \leq 0.360$. The additive disturbance, d, was uniformly distributed between -0.8 and $+0.8$, with the true value of a given by $a = \underline{a} + d$. With each branch, i.e., each $y_i = i$, there is a probability, p_i, of a penalty for $I.P. > \theta$. As θ decreases, each p_i increases until only one, and finally no, $p_i < 0.5$. As the automaton adapts (learns), it should eventually become imbedded in the branch with $p_i = M_{min}$.

D. SIMULATION RESULTS

The results using the linear tactic automaton are shown in Figures 4, 5, and 6 where the average values of I.P. and penalty are plotted as a function of the number of states in each branch. The averages were taken over 3800 transitions for case 1 and 2000 transitions for cases 2 and 3.

1. Case 1 -- All Branches with $p_i > 0.5$

This case, shown in Figure 4, occurs when θ is near the minimum possible I.P. The linear tactic model is not optimum, but the Krylov modification yields definitely improved results with the average I.P. and penalty decreasing as n increases. There is a tendency to get caught in a branch having $p_i > p_{min}$ with Krylov's model, however, since the effective p_i in all branches is no greater than 0.5. As θ is decreased, and p_{min} increased, all but one branch have $p_i = 1.0$ (effective $p_i = 0.5$) and this tendency is reduced.

2. Case 2 -- Only One Branch with $p_i < 0.5$

As seen in Figure 5, the linear tactic model gives excellent results in this case. As n increases the average penalty and I.P. decrease, but for relatively small values of n (e.g., 8), the automaton behaves as if n were very large. Thus it may not be necessary or practical to allow n to approach infinity. The Krylov model gives poor performance in this case since it effectively reduces the p_i in more than one branch to less than 0.5.

3. Case 3 -- More than One Branch with $p_i < 0.5$

This case, shown in Figure 6, occurs when θ is relatively high. The automaton may become imbedded in any branch with $p_i < 0.5$. The worst case may always be investigated by initially placing the automaton in state s_{41}, where $p_3 < p_4 < 0.5$. Computer data shows that the automaton tended to remain in branch 4, sometimes never leaving that branch in 2000 transitions, even though $p_3 = p_{min} = 0.0 < p_4$. As n increased, the tendency to remain in any branch with $p_i < 0.5$ increased. Only the linear tactic results are shown for this case.

First results indicate the Varshavskii and Vorontsova model behaves in a manner closely following predictions. It appears not to be dependent on whether $p_{min} < 0.5$. The choice of adjustment parameter and the reinforcement method have not been investigated sufficiently to justify stating any definite results at this time, however.

The choice of θ determines the various p_i. If θ is too large, more than one branch has $p_i < 0.5$, but if θ is too low, all $p_i > 0.5$ and the automaton tends to circulate around all branches. A possible algorithm for determining the best value of θ is the following: allow the automaton to reach a given depth in any branch. Assume this branch has $p_i < 0.5$ and change branches. Determine which other branches have $p_i < 0.5$ for the current θ. Eliminate branches with $p_i > 0.5$, lower θ, and repeat until only one branch remains.

Observation of Figures 4-6 shows that the I.P. and average penalty have similar behavior, even though no attempt has been made to relate I.P. and M directly. In the following sections, a method is developed for choosing Θ and the quantization width, and the automaton model is modified.

III. ADAPTIVE MODEL

A. TANGENT-SLOPE METHOD

Assume that the I.P. is a continuous function of a single parameter: I.P. $= y(z)$. This function may be multi-modal and is assumed to have v minima in the admissible region, $0 \leq z \leq Z$. The "tangent-slopes", m_{ij}, $(i = 1, \ldots, v\ ;\ j = 1, 2)$ are defined as the slopes of the lines from each minimum which are tangent to the I.P. curve, but nowhere cutting the curve above the respective minima. In the admissible region define the maximum tangent-slope as

$$\mu = \max |m_{ij}| \quad (i = 1, \ldots, v\ ;\ j = 1, 2) \tag{14}$$

Typical tangent-slopes are illustrated in Figure 7.

For a given Θ, it is desired to find a maximum quantization width, W, of the admissible region in order to eliminate the largest area possible from further consideration, but not eliminate the area, W_o, containing the true minimum, $y_o = y(z_o)$. Using the automata described, at least half of the curve in W_o must lie below Θ, i.e., $p_i < 0.5$. Observe that m_{ij} effectively defines an envelope inside of which y must lie.

Next consider a worst case where on both sides of z_o the $|m_{ij}| = \mu$, as shown in Figure 8. Let W have one end at z_o and the other at z_x, where $y(z_x) \leq 2\Theta - y_o$. Then for $p_i \leq 0.5$, $\mu \leq \dfrac{2(\Theta - y_o)}{W}$.

Rearranging yields

$$W \leq \frac{2(\Theta - y_o)}{\mu} = W_{max} \quad \text{and } \Theta \geq \frac{W}{2} \mu + y_o = \Theta_{min}. \qquad (15)$$

If $|m_{i1}|$ and $|m_{i2}|$ at z_o are less than the estimated or known μ , as suggested by the dotted lines, m'_{o1} and m'_{o2} , in Figure 8, (15) guarantees that $p_i < 0.5$ in the area containing y_o if z is varied uniformly over each area, W.

Equation (15) assumes knowledge of μ and y_o. In the usual practical situation, the value of μ must be estimated if the exact curve is not known. This may serve as a measure of the "roughness" of the curve - if the curve is relatively smooth over the entire admissible region, then μ has a relatively low value. If not known, y_o is estimated by recording the minimum I.P. value measured as the operation progresses. Then this minimum value is the estimated minimum, $\bar{y}_o \geq y_o$. If $\bar{y}_o$ is used in (15), Θ_{min} and W_{max} may be conservatively determined. As the operation continues, $\bar{y}_o$ should be reduced.

Define the wavelength, r , as the distance between two local minima. Then define an average, $\bar{r} = \frac{\sum_{i=1}^{v-1} r_i}{v - 1}$. As an estimate of $\bar{r}$, define $\hat{r} = \frac{Z}{v}$. The initial value of W is an arbitrary choice, W(0). A large W(0) gives larger areas of investigation and a possibility of eliminating large areas early, but by (15) also yields a high Θ_{min}. As a reasonable initial choice, let

$$W(0) = \frac{\hat{r}}{4} = \frac{Z}{4v} \quad . \qquad (16)$$

The above development guarantees that if each area, W, is represented by a branch of the automaton, the area containing the true minimum has $p_i < 0.5$. From (16) the admissible region is initially divided into $k_o = 4v$ equal areas, yielding $4v$ branches in the automaton.

B. MODIFIED AUTOMATON

A modification of the previous model is shown in Figure 9. Assume the automaton is initially in the middle state of branch i. If $x(t) = 1$ (penalty), the automaton moves sequentially toward s_{i1}, and if $x(t) = 0$, the transition is toward s_{in}, where n is odd. If s_{i1} or s_{in} is reached, branch i is eliminated or retained, respectively. After either s_{i1} or s_{in}, the next state is $s_{i+1,\frac{n+1}{2}}$, unless all branches have been investigated, in which case the areas, W, are halved and all remaining branches are replaced by a pair of branches. θ is lowered in accordance with (15).

C. MARKOV ANALYSIS

Consider any single branch of this automaton. If s_{i1} and s_{in} are considered as absorbing states, i.e., no other state in branch i can be reached from either state, the transition matrix becomes:

$$P = \begin{array}{c}
\begin{array}{ccccccccccc}
s_{i1} & s_{i2} & s_{i3} & \cdots & s_{ij-1} & s_{ij} & s_{i,j+1} & \cdots & s_{i,n-2} & s_{i,n-1} & s_{in}
\end{array} \\
\left[\begin{array}{ccccccccccc}
1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
q_i & 0 & p_i & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & q_i & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & p_i & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & q_i & 0 & p_i & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & q_i & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & p_i & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & q_i & 0 & p_i \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1
\end{array}\right]
\end{array} \tag{17}$$

Thus P is a random walk with absorbing barriers, and each branch may be analyzed as the classical "gambler's ruin" problem.(8) Let π_z be the probability of reaching s_{i1} and γ_z the probability of reaching s_{in}, starting in state $s_{i,\frac{n+1}{2}}$. It can be shown(9) that

$$\pi_z = \frac{1}{1 + \left(\frac{q_i}{p_i}\right)^{\frac{n-1}{2}}} = 1 - \gamma_z \qquad (18)$$

Since $\pi_z + \gamma_z = 1$, either state s_{i1} or s_{in} must eventually be reached. If D_z is the expected number of transitions to reach either s_{i1} or s_{in}, then

$$D_z = \frac{n-1}{p_i - q_i}\left[\frac{1}{2} - \gamma_z\right] \qquad p_i \neq q_i\ , \ n > 1 \qquad (19)$$

$$D_z = \frac{1}{4}(n-1)^2 \qquad p_i = q_i = 0.5\ , \ n > 1 \qquad (20)$$

Figure 10 is a plot of π_z and D_z versus the number of states, n, in a branch for various $p_i < 0.5$. It is desirable that π_z approach zero ($\gamma_z \rightarrow 1$) so that branch i is not eliminated. As n increases, π_z decreases, but D_z increases. Thus some compromise is required to keep the search time practical and yet minimize the probability of eliminating an area with $p_i < 0.5$.

D. ILLUSTRATIVE EXAMPLES

The modified automaton, with $n = 9$, was used to determine the absolute minimum of fourth and sixth order I.P. curves with two and three minima, respectively ($v = 2, 3$). The curves are described by

$$\text{I.P.} = y = 3.125z^4 - 23.75z^3 + 56.875z^2 - 46.25z + 20.0 \qquad (21)$$

$$\text{I.P.} = y = \frac{1}{240} z^6 - \frac{5.4}{40} z^5 + \frac{269}{160} z^4 - \frac{1213}{120} z^3 + \frac{1185}{40} z^2 - 35_z + 17 \qquad (22)$$

The curve of (21) is shown in figure 11 (a) where the absolute minimum occurs at $z_o = 3.25$, $y_o = 3.8$ and $\mu = 28.3$. The admissible region is $0 \leq z \leq 4 = Z$, and from (16), $W(0) = 0.5$ and $k_o = 4v = 8$. Figure 11 shows the elimination of portions of the curve as the operation progresses. After 140 transitions and three modifications (halving of W), the local minimum at $z = 0.6$ was eliminated and an area containing the true minimum remained. The automaton investigated and reduced this area until, after 314 transitions, the remaining area was $3.1562 \leq z \leq 3.3437$, a total width of $W_T = 0.1875$. The number of branches in the automaton had increased to $k = 24$ and Θ was reduced to 3.892. In Figure 11, when the plotted value of Θ is considered, it appears that more areas should be eliminated. However, Θ is continually lowered during the operation and some areas were investigated for higher Θ than the one shown, i.e., the value of Θ at the completion of the indicated number of transitions.

Figure 12 (a) shows the curve of (22) with an absolute minimum at $z_o = 10$, $y_o = 0.33$. The admissible region is $0 \leq z \leq 12$ and $\mu = 53.57$. However, μ was only estimated to be 15.0, and $W(0) = 1.0$ and $k_o = 4v = 12$. After 598 transitions, $k = 38$, $\Theta = 0.363$, and the remaining area was $9.9219 \leq z \leq 10.0703$, a total width of $W_T = 0.1484$. Although the estimated value of μ was obviously too low, the automaton did not eliminate the true minimum in 20 experiments.

Figure 13 is a plot of k (number of branches), average I.P., Θ, and W_T (total remaining area) versus the number of transitions for (22).

The average I.P. is the average value over the period of transitions between automaton modifications. Each point plotted represents a time of automaton modification. As the operation progressed, the number of branches became very large, thus requiring longer periods of investigation, and the average I.P. decreased very slowly. No attempt was made in these experiments to determine an algorithm for ending the experiment.

IV. CONCLUSIONS

The Tsetlin model of automata, operating in a random environment, offers several advantages of a practical nature. The entire system is described by a rather simple Markov chain, thus providing a model for thorough mathematical analysis. The model is easily modified and it is shown that one possible application is in control systems where the adaptive behavior makes the automaton suitable for use as a learning controller. Other systems of possible application might be pattern recognition devices and communication systems.

The results reported in this paper are very satisfactory and in accord with the theory developed. In the particular application reported here, several problems remain to be solved, e.g., algorithms for terminating a search for optimum, introduction of memory when different environments are recognizable, and algorithms for adjusting θ until only one branch remains with $p_i < 0.5$. The probabilistic model of Varshavskii and Vorontsova is currently under further investigation. The learning properties of this model appear to have interesting and promising possibilities for application.

REFERENCES

(1) M. L. Tsetlin, "On the Behavior of Finite Automata in Random Media," Automatika i Telemekhanika, Vol. 22, No. 10 (October, 1961), pp. 1345-1354.

(2) V. I. Varshavskii and I. P. Vorontsova, "On the Behavior of Stochastic Automata with a Variable Structure," Automatika i Telemekhanika, Vol. 24, No. 3, (March, 1963), pp. 353-360.

(3) J. E. Gibson, K. S. Fu, et.al., "Philosophy and State of the Art of Learning Control Systems," Report TR-EE 63-7, Control and Information Systems Laboratory, School of Electrical Engineering, Purdue University, AFOSR-5144 (November, 1963).

(4) K. S. Fu, "Learning Control Systems," COINS Symposium, June 17-18, 1963, Evanston, Illinois.

(5) J. D. Hill, G. J. McMurtry, and K. S. Fu, "A Computer-Simulated On-Line Experiment in Learning Control Systems," AFIPS Proceedings, 1964 Spring Joint Computer Conference, Vol. 25, pp. 315-325.

(6) R. R. Bush and F. Mosteller, Stochastic Models for Learning, New York: John Wiley and Sons, 1955.

(7) V. Y. Krylov, "On One Automaton that Is Asymptotically Optimal in a Random Medium," Automatika i Telemekhanika, Vol. 24, No. 9 (September, 1963), pp. 1226-1228.

(8) W. Feller, An Introduction to Probability Theory and its Applications, New York: John Wiley and Sons, 1957.

(9) G. J. McMurtry, "A Probabilistic Automaton as a Learning Controller," Ph.D. Thesis in preparation, Lafayette, Indiana: School of Electrical Engineering, Purdue University.

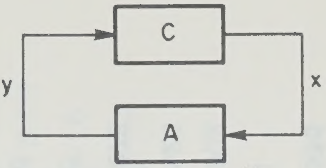


Fig. 1. Automaton, A, Operating in a Random Environment, C

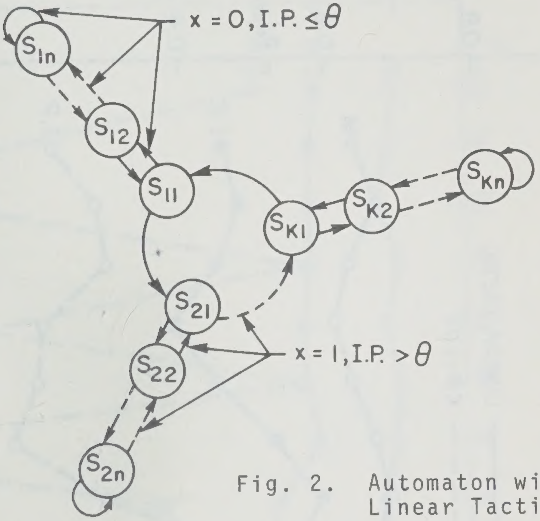

x = 0, I.P. ≤ θ
S_{1n}
S_{12}
S_{11}
S_{Kn}
S_{K2}
S_{K1}
S_{21}
S_{22}
x = 1, I.P. > θ
S_{2n}

Fig. 2. Automaton with Linear Tactic

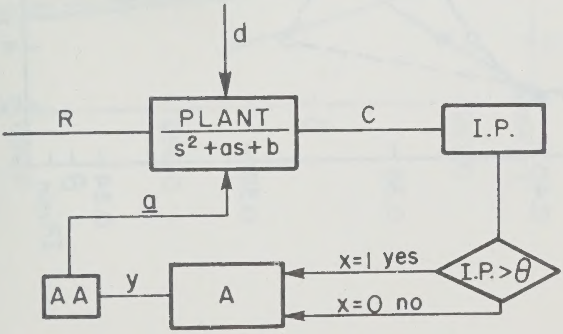

d
R
PLANT
$s^2 + as + b$
C
I.P.
a
AA
y
A
x=1 yes
I.P.>θ
x=0 no

Fig. 3. Learning Control System

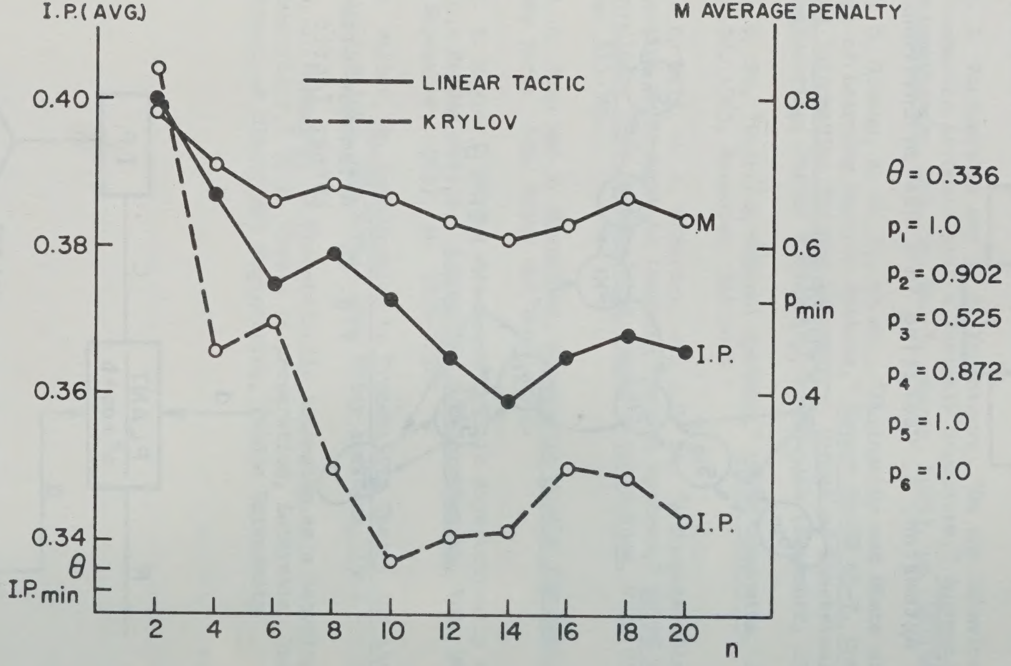

I.P.(AVG.)
M AVERAGE PENALTY
LINEAR TACTIC
KRYLOV
0.40
0.38
0.36
0.34
θ
I.P. min
0.8
0.6
p min
0.4
2
4
6
8
10
12
14
16
18
20
n
M
I.P.
I.P.
θ = 0.336
p1 = 1.0
p2 = 0.902
p3 = 0.525
p4 = 0.872
p5 = 1.0
p6 = 1.0

Fig. 4. Case 1

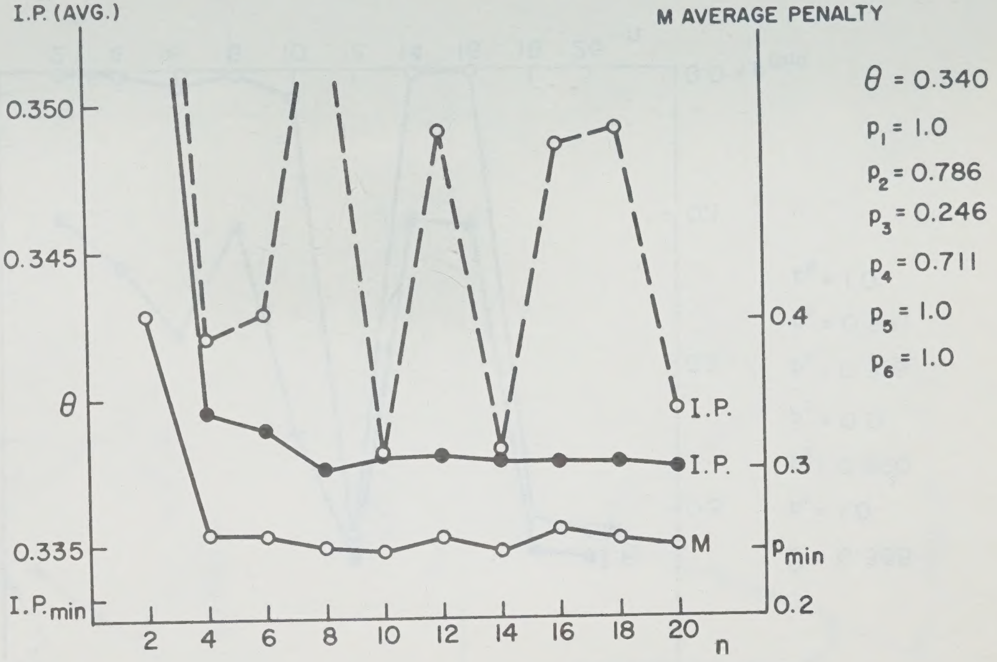

I.P. (AVG.)
M AVERAGE PENALTY
0.350
0.345
θ
0.335
I.P.$_{min}$
2
4
6
8
10
12
14
16
18
20
n
I.P.
I.P.
M
0.4
0.3
p_{min}
0.2
θ = 0.340
p_1 = 1.0
p_2 = 0.786
p_3 = 0.246
p_4 = 0.711
p_5 = 1.0
p_6 = 1.0

Fig. 5. Case 2

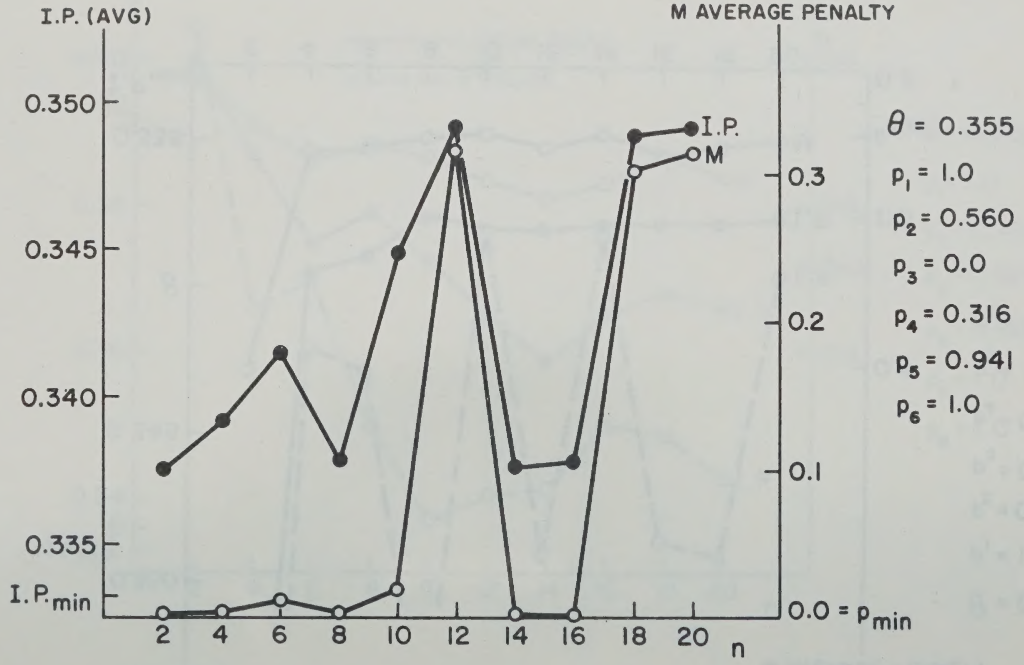

I.P. (AVG)
M AVERAGE PENALTY
0.350
0.345
0.340
0.335
I.P.$_{min}$
0.3
0.2
0.1
0.0 = p$_{min}$
2
4
6
8
10
12
14
16
18
20
n
I.P.
M
θ = 0.355
p$_1$ = 1.0
p$_2$ = 0.560
p$_3$ = 0.0
p$_4$ = 0.316
p$_5$ = 0.941
p$_6$ = 1.0

Fig. 6. Case 3

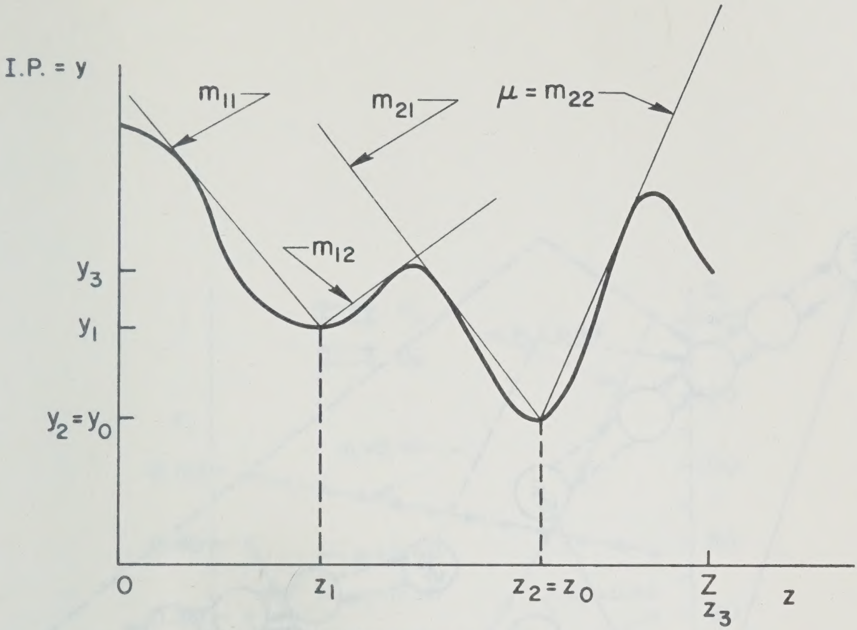

I.P. = y
m_{11}
m_{21}
$\mu = m_{22}$
m_{12}
y_3
y_1
$y_2 = y_0$
0
z_1
$z_2 = z_0$
Z
z_3
z

Fig. 7. Tangent-slopes

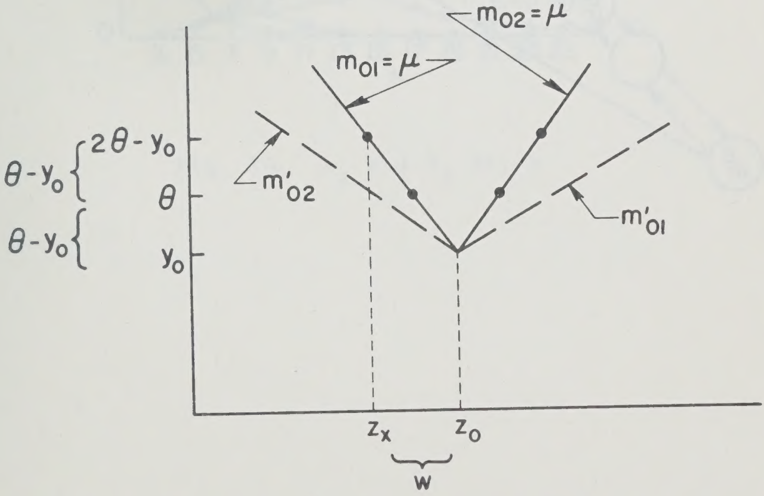

$m_{02} = \mu$
$m_{01} = \mu$
$2\theta - y_0$
$\theta - y_0$
θ
$\theta - y_0$
y_0
m'_{02}
m'_{01}
z_x
z_0
w

Fig. 8. Worst Case for Finding θ_{min}, W_{max}

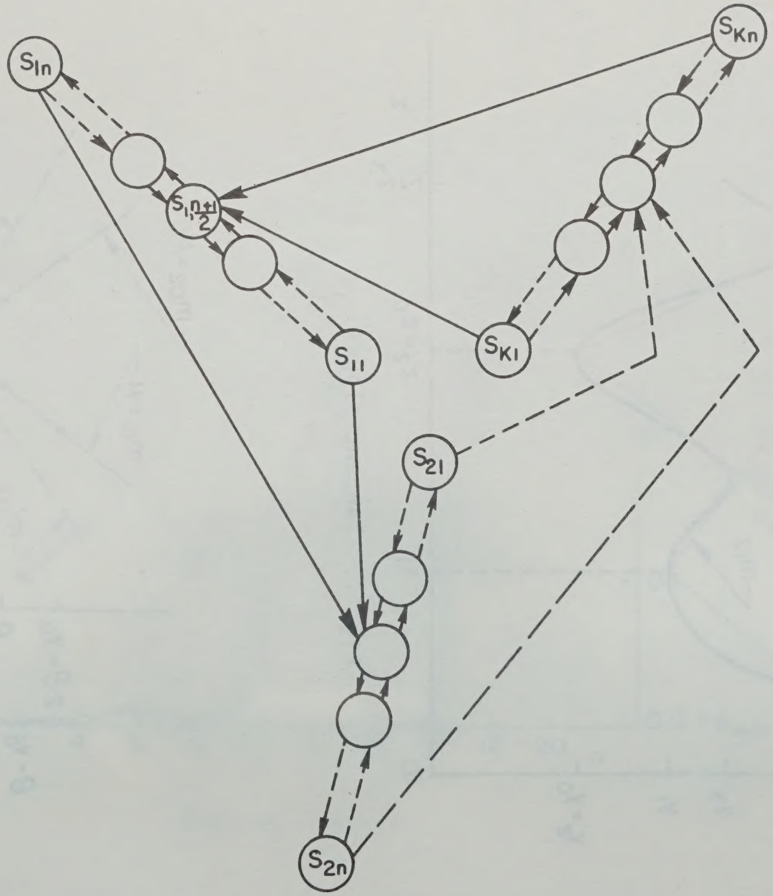

S_{1n}
$S_{1,\frac{n+1}{2}}$
S_{11}
S_{Kn}
S_{K1}
S_{21}
S_{2n}

Fig. 9. Modified Automaton

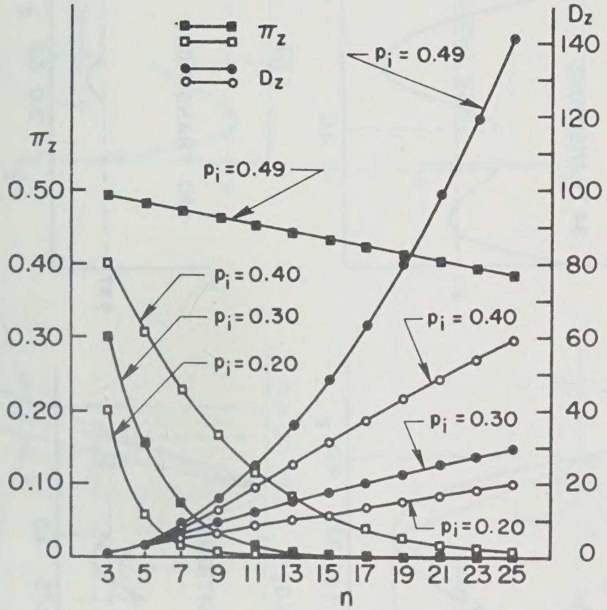


Fig. 10. π_Z and D_Z vs. n

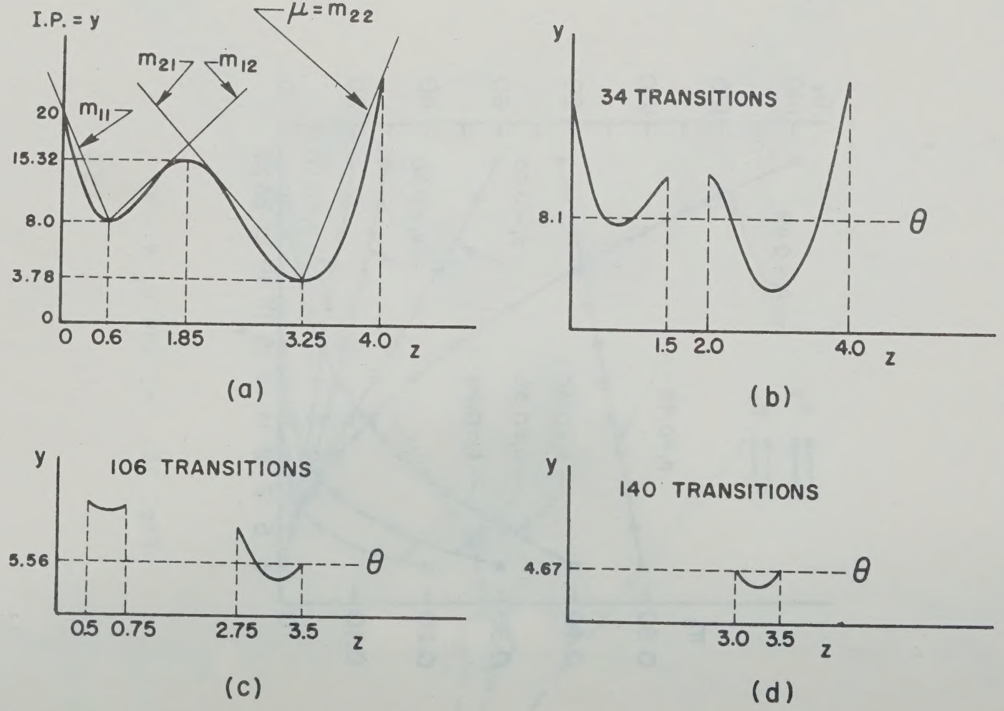


Fig. 11. Results for 4th-order I. P.

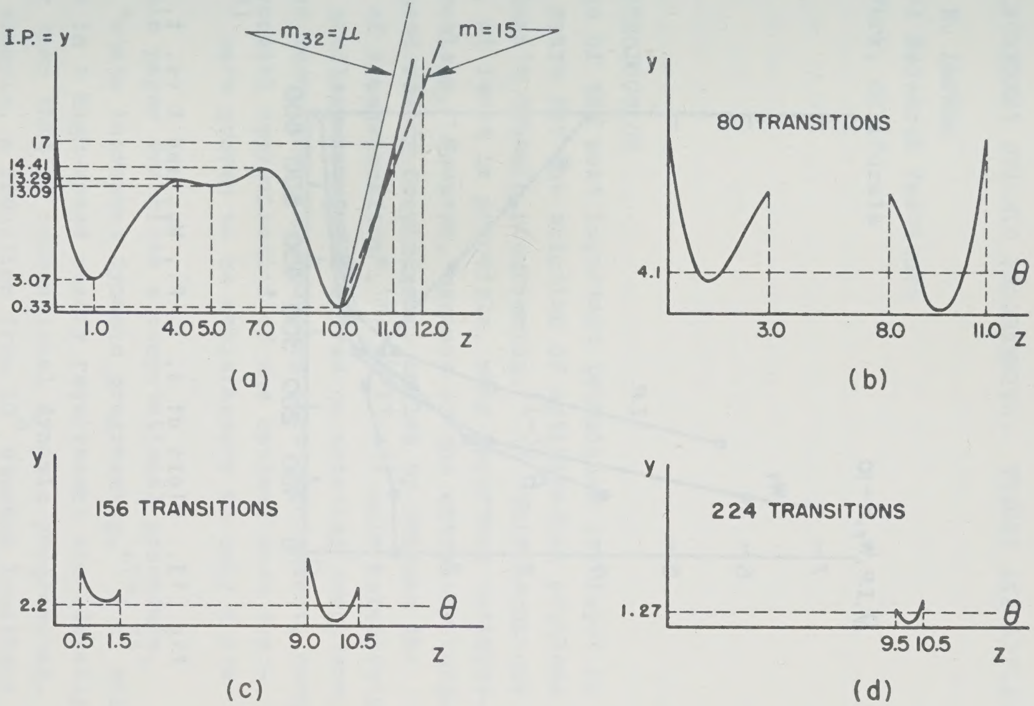


Fig. 12. Results for 6^{th}-order I. P.

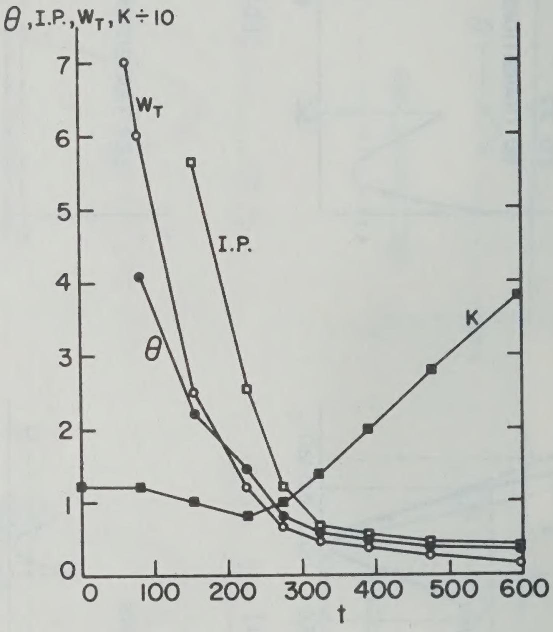


Fig. 13. Plots of θ, I. P., W_T, and K vs. t.

STATE INCREMENT DYNAMIC PROGRAMMING: THEORY AND APPLICATIONS

Robert E. Larson
Stanford Research Institute
Menlo Park, California

I INTRODUCTION

One of the most important techniques developed in recent years for the solution of optimization problems is Bellman's dynamic programming.(1-3) This technique solves, at least in principle, many important optimization problems. However, because of the extremely large high-speed memory requirement, called by Bellman the "curse of dimensionality",(2, Chapt. 5) only relatively simple problems have been solved on existing computers. Proposed techniques for reducing this requirement, such as polynomial approximation of the optimal cost function,(3) have proven to be satisfactory in only a few cases.

This paper describes a computational procedure, called "state increment dynamic programming,"(4, 5) which results in a high-speed memory requirement considerably smaller than that of conventional dynamic programming. In one example, a reduction from 10^6 storage locations to 100 locations is obtained. A simultaneous decrease in computing time is often achieved. This procedure is applicable to most of the problems in which the conventional method can be used.

The new procedure, like the conventional method, is based on iterative application of Bellman's principle of optimality. It differs from the conventional method in the choice of the time interval over which a given control is applied. Instead of using a fixed interval, the new procedure determines the time interval as the minimum time required for at least one of the state variables to change by one increment. As a result of this choice of interval, the next state for any given control is known to lie within a small neighborhood of the point at which control is applied. By using this result it is possible to compute optimal control in units, called blocks, that cover a relatively long time interval but a small distance along each state variable. By using only one or two high-speed memory locations per state in the block--an almost arbitrarily small number of locations--it is possible to compute optimal control throughout the block. Specialized computations near the boundaries of the block allow the optimal trajectories any desired degree of freedom in passing from block to block.

This technique has been applied in a computer program which calculates minimum fuel trajectories for supersonic (Mach 3) aircraft under a variety of conditions and constraints. The basic program can be used both for a detailed evaluation of an aircraft design and for real-time control of an aircraft. Computation of the optimal bounded control of linear plants has also been studied in detail, and a number of other possible applications are indicated.

II PROBLEM DESCRIPTION

Although the method can be extended to cover other cases, in this paper the system equations will be considered to be a set of nonlinear time-varying differential equations subject to a number of constraints. The system equations are

$$\underline{\dot{x}} = \underline{f}\ (\underline{x},\ \underline{u},\ t) \tag{1}$$

where

$\underline{x}$ = n-dimensional state vector

$\underline{u}$ = q-dimensional control vector

t = independent variable, usually time

$\underline{f}$ = n-dimensional vector functional

$(\dot{\ })$ = d/dt ( ).

The constraints can be of the type permitted in conventional dynamic programming, such as multiple time-varying inequality constraints on both control and state variables.

In order to implement the procedure on a digital computer, the set of differential equations is replaced by a set of difference equations.

$$\underline{x}\ (t + \delta t) = \underline{x}\ (t) + \underline{f}\ (\underline{x}(t),\ \underline{u},\ t)\ \delta t \qquad (2)$$

where δt is an incremental time.

The performance criterion, which is to be minimized, is taken to be the integral over fixed time limits of a scalar function of state variables, control variables, and the independent variable.

$$J\ (\underline{x},\ \underline{u},\ t_0) = \int_{t_0}^{t_f} \ell\,(\underline{x},\ \underline{u},\ \sigma)\ d\sigma \qquad (3)$$

where

J = cost function

ℓ = penalty function; cost function per unit time

σ = dummy variable for time

t_0 = initial time

t_f = final time.

Again, the performance criterion can easily be extended to cover many other cases.(3, 6)

The problem can be stated as follows: For all admissible states, $\underline{x} \in X$, and for all t, $t_0 \leq t \leq t_f$, compute that control function from the set of admissible control functions, U, which minimizes J $(\underline{x},\ \underline{u},\ t)$. As in conventional dynamic programming, the minimum cost function is defined as I $(\underline{x},\ t)$.

$$\begin{aligned} I\ (\underline{x},\ t) &= \min_{\underline{u}\ \in\ U} \left\{ J\ (\underline{x},\ \underline{u},\ t) \right\} \\ &= \min_{\substack{\underline{u}(\sigma)\ U \\ \underline{t \leq \sigma \leq t_f}}} \left\{ \int_{t}^{t_f} (\underline{x},\ \underline{u},\ \sigma)\ d\sigma \right\} \qquad (4) \end{aligned}$$

The optimal control function at state $\underline{x}$, time t, is denoted as $\hat{\underline{u}}(\underline{x}, t)$.

The new procedure, like conventional dynamic programming, solves this problem by iterative application of Bellman's principle of optimality.$^{(1-3)}$ For the problem defined in this section, the principle of optimality has the following form:

$$I(\underline{x}, t) = \min_{\underline{u}(t)\in U} \Big\{ \ell(\underline{x}, \underline{u}(t), t)\ \delta t + I\left[\underline{x} + \underline{f}(\underline{x}, \underline{u}(t), t)\ \delta t,\ t + \delta t\right] \Big\} \tag{5}$$

III DIFFERENCE FROM THE CONVENTIONAL METHOD

The sets of admissible states and admissible controls are determined exactly as in conventional dynamic programming. Contraints in the problem are used to restrict each state variable to a finite range.

$$\beta_i^- \leq x_i \leq \beta_i^+ ; \quad i = 1, 2, \ldots, n. \tag{6}$$

Within this range each state variable is quantized in uniform increments Δx_i.

$$\begin{aligned} x_i &= \beta_i^- + j_i\ \Delta x_i \\ j_i &= 0, 1, \ldots, N_i \\ N_i\ \Delta x_i &= \beta_i^+ - \beta_i^- \\ i &= 1, 2, \ldots, n. \end{aligned} \tag{7}$$

The set of admissible states, X, consists of vectors $\underline{x}$, where each component x_i is of the form of Eq. (7). The set can lie in any convex region of the total state space and can vary with time.

The set of admissible controls can be determined similarly. It is sufficient to state that the set of admissible controls U contains a finite number of elements.

$$U = \left\{\underline{u}^{(1)}, \underline{u}^{(2)}, \ldots, \underline{u}^{(K)}\right\} \tag{8}$$

The basic difference between the new procedure and conventional dynamic programming is in the method of determining the time interval over which a given control is applied. With the conventional method, this interval is a fixed value Δt.

$$\delta t = \Delta t. \tag{9}$$

In the new procedure, on the other hand, this time interval is determined as the interval necessary for at least one of the n state variables to change by one increment.

$$\delta t = \min_{i = 1, 2, \ldots, n} \left\{ \left| \frac{\Delta x_i}{f_i(\underline{x}, \underline{u}, t)} \right| \right\} \tag{10}$$

The fact that control is applied until a state changes by one increment rather than for a fixed time increment is the source of the term "state increment dynamic programming."

An important consequence of this choice of δt is that the next state always lies within a small neighborhood of the present state. Specifically, the next state lies on the surface of an n-dimensional hypercube centered at the present state and with length $2\Delta x_i$ along the x_i-axis. The new procedure uses this property to reduce the high-speed memory requirement from that of the conventional method.

IV COMPUTATIONAL PROCEDURE

A. INITIALIZATION PROCEDURE

In order to begin the computational procedure, it is necessary to have available one value of the optimal cost function for each quantized state, $\underline{x} \in X$. It is not necessary that each value be computed at the final time t_f as in conventional dynamic programming. However, if the value for a given state is computed at t', then, for $t > t'$, no computation of optimal control will be performed at that state.

An especially convenient choice of t' for each state can be made when the terminal state $\underline{x}(t_f)$ is constrained to fall within C, a subset of X. In this case $t' = t_f$ if $\underline{x} \in C$, while if $\underline{x}$ is not in C, t' is taken to be the least time for which it is still possible to reach C at t_f from the given state using admissible control.

B. DEFINITION OF THE BLOCK

The first step in the reduction of the high-speed memory requirement is the partitioning of the (n + 1)-dimensional X-t space into rectangular units called blocks. Each block covers w_i increments along the x_i-axis and ΔT seconds along the t-axis. A given block is denoted by the largest values of the coordinates that are contained in the block.

$$B\ (j_1,\ j_2,\ \ldots,\ j_n,\ m) = \underline{x},\ t \tag{11}$$

such that

$$(j_i - 1)\ w_i\ \Delta x_i \leq x_i \leq j_i\ w_i\ \Delta x_i$$

$$(m - 1)\ \Delta T \leq t - t_0 \leq m\ \Delta T$$

where

$$j_i = 1,\ 2,\ \ldots,\ J_i$$

$$m = 1,\ 2,\ \ldots,\ M$$

and

$$J_i\ w_i\ \Delta x_i = \beta_i^+ - \beta_i^-;\quad i = 1,\ 2,\ \ldots,\ n.$$

$$M\ \Delta T = t_f - t_0.$$

As indicated in the above equation, the boundary between adjacent blocks is considered to be included in both blocks. The number w_i is taken to be a small integer, usually between 2 and 5; and ΔT is taken to be considerably larger than the average value of δt determined by Eq. (10).

C. COMPUTATIONS WITHIN A BLOCK

The next step in the reduction of the high-speed memory requirement is to constrain the next state for any control to be in the same block as the state at which the control is applied. Since from the choice of t, the next state must always lie in a neighborhood of the present state, this constraint is easy to implement. The reduction in high-speed memory requirement occurs because only one or two values of the optimal cost function per state in the block are needed in the high-speed memory, rather than one value per state in the entire state space as in the conventional method.

The remainder of this section is devoted to a detailed description of the computational procedure when the state at which the computations are taking place is inside the block. In this case it is known that the next state must be in the same block. The modifications necessary when the state lies on the boundary of the block and how these modifications permit interblock transitions to take place is discussed in the next section.

1. Procedure in the General Case

In the general case, the time interval ΔT is further divided into smaller increments, ΔT. A set of quantized times is determined.

$$t = t_0 + (m - 1)\,\Delta t + s\,\Delta t \tag{12}$$

where

$$s = 0, 1, \ldots, S$$

$$S\,\Delta t = \Delta T.$$

Typical values of S range from 5 to 15. Optimal control is computed at each quantized state $\underline{x} \in X$ for each quantized t of the form in Eq. (12). The operations at a given $\underline{x}$ and t take place as follows.

Each admissible control, $\underline{u}^{(k)} \in U$, is applied. For each control, the time over which the control is applied is determined as $\delta t^{(k)}$ from Eq. (10).

$$\delta t^{(k)} = \min_{i = 1, 2, \ldots, n} \left\{ \left| \frac{\Delta x_i}{f_i\,(\underline{x}, \underline{u}^{(k)}, t)} \right| \right\} \tag{13}$$

The corresponding next state is computed.

$$\underline{x}^{(k)} = \underline{x} + \underline{f}\,(\underline{x}, \underline{u}^{(k)}, t)\;\delta t^{(k)}. \tag{14}$$

The optimal cost function at this state and time, $I\,(\underline{x}^{(k)}, t + \delta t^{(k)})$, is computed by interpolation in (n-1) state variables and time using the previously computed values at quantized states and at times $t + \Delta t$, $t + 2\Delta t$, ... If the increment Δt is chosen to be of the same order of magnitude as the average value of δt, then linear interpolation in time based on values at $t + \Delta t$ and $t + 2\Delta t$ is generally sufficient. The state variable not used in the interpolation formula is the one for which $\delta t^{(k)}$ in Eq. (13) takes on the minimum value.

If a control $\underline{u}^{(k)}$ is such that none of the state variables change, i.e., if $\underline{f}(\underline{x}, \underline{u}^{(k)}, t) = 0$, then $\underline{x}^{(k)} = \underline{x}$ and $\delta t^{(k)}$ is set equal to Δt.

The resulting next states for a one-dimensional example are shown in Figure 1, where $U = \{\underline{u}^{(1)}, u^{(2)}, \underline{u}^{(3)}, \underline{u}^{(4)}\}$. Values of the optimal cost function are known at the points indicated by open circles.

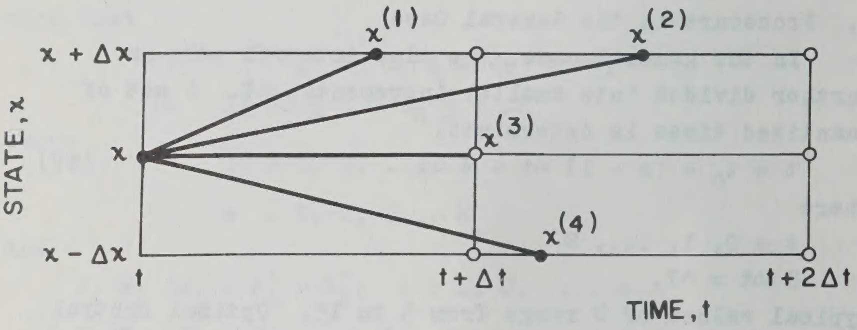


Figure 1: Procedure in the General Case for a One-Dimensional Example

Once the optimal cost for each next state has been found, the cost of each control over the interval $\delta t^{(k)}$ is found by evaluating $\ell(\underline{x}, \underline{u}^{(k)}, t)$, the cost function per unit time. The optimal cost and optimal control at $\underline{x}$, t are then calculated from the principle of optimality.

$$I(\underline{x}, t) = \min_{k = 1, 2, \ldots, K} \{\ell(\underline{x}, \underline{u}^{(k)}, t)\, \delta t^{(k)} + I(\underline{x}^{(k)}, t + \delta t^{(k)})\} \quad (15)$$

$$\hat{k} = \arg \min_{k = 1, 2, \ldots, K} \{\ell(\underline{x}, \underline{u}^{(k)}, t)\, \delta t^{(k)} + I(\underline{x}^{(k)}, t + \delta t^{(k)})\} \quad (16)$$

$$\hat{\underline{u}}(\underline{x}, t) = \underline{u}^{(\hat{k})} \quad (17)$$

2. Procedure with $q \leq n$

When the set of admissible controls is such that each admissible control changes at most one state variable, the above procedure can be modified. The most important case where this situation exists is when q, the number of control variables, is greater than or equal to n, the number of state variables. When these modifications are made, interpolations take place in one dimension, time, rather than in n dimensions. There is also a significant reduction in both the high-speed memory requirement and computing time. The modified procedure is described in detail in references 4 and 5.

D. INTERBLOCK TRANSITIONS

In order to allow optimal trajectories to pass from block to block, the computations near the boundaries of the block must be slightly modified. The modifications are similar for both of the previously described procedures, but a different modification is used, depending on whether or not the adjacent block at the boundary has been processed.

For two adjacent blocks, one of which has been processed and the other has not, a sequence of optimal points has been computed along the boundary between these blocks. These optimal points are stored in the high-speed memory. No new optimal points are computed along this boundary; however, the values of the optimal cost function for these points are used in evaluating the cost of these controls applied at states within one increment of the boundary for which the next state lies on this boundary. It is then true that an optimal trajectory can pass from a state on the interior of the block to the boundary with a previously computed block. From this boundary the optimal control may take the trajectory further into the previously computed block. In this manner, transitions from a given block to an adjacent block that has been previously processed can take place without constraint, provided that optimal points on the boundary are available.

In most problems of physical significance it is possible to obtain enough feeling about the system to determine the direction which optimal trajectories are most likely to take. This direction is called the preferred direction of motion. In many linear problems it is known that the preferred direction is toward the origin. In the supersonic aircraft control problem discussed in a later section, it is known that as the aircraft consumes fuel and becomes lighter, it should fly at higher altitudes and higher velocities.

When a preferred direction of motion can be defined, adjacent blocks in a given time interval are processed in an order such that the preferred direction is from the block processed later to the one processed first. In this manner, interblock transitions in the preferred direction take place without constraint.

Although interblock transitions in the opposite direction will occur less frequently, they must nevertheless be considered. Several procedures are available for implementing these transitions. First, after two adjacent blocks in the same time interval have been processed, it is possible to compute interblock transitions in the nonpreferred direction at $t = (m - 1) \Delta T$, the least time in the block. These transitions can be compared with the previously computed optimal points on the boundary at this time. This procedure has the affect of constraining the rate of change in a nonpreferred direction to be less than $(w_i \Delta x_i)/\Delta T$. The quantity ΔT can sometimes be chosen so that this constraint is not unduly restrictive, and yet a reasonable number of computations per state are obtained in the interval.

If more frequent interblock transitions in the nonpreferred direction are desired, it is possible to use values of the optimal cost within the block being computed to extrapolate the optimal cost function into a not-yet-computed block. Since the extrapolation is limited to extend at most Δx_i along any coordinate x_i, the errors introduced by this technique can be bounded. The computations described in the previous paragraph can be used to provide more accurate results every Δt seconds and hence limit the accumulation of errors from the extrapolation process.

In addition, recomputation of optimal points along the boundary between blocks eliminates much of the inaccuracy introduced in these transitions. However, the extra computing time required may be excessive in terms of the improvement obtained, especially if a combination of the other two methods yields satisfactory results.

E. HIGH-SPEED MEMORY REQUIREMENT

For the conventional dynamic programming procedure, if there are n state variables and N_i quantization levels in the ith state variable, the high-speed memory requirement is given as follows:(1, 3)

$$N = \prod_{i=1}^{n} N_i \tag{18}$$

For a three-dimensional system with 100 quantization levels in each state

$$N = (100)^3 = 10^6.$$

In the new procedure, however, a much lower number is required. If a block covers $w_i \; \Delta x_i$ units along the ith axis, so that there are $(w_i + 1)$ quantization levels in each block, then for the case where $q \geq n$, the high-speed memory requirement is

$$N' = \prod_{i=1}^{n} (w_i + 1) \tag{19}$$

For the general case using linear interpolation in time, twice as many locations (2N') are required. If, in addition, points on the boundaries with previously computed blocks are stored in the high-speed memory, and if there are S optimal points per state in an interval Δt, then the high-speed memory requirement becomes

$$N'' = 2\prod_{i=1}^{n} w_i + S\left(\prod_{i=1}^{n} (w_i + 1) - \prod_{i=1}^{n} w_i\right). \quad (20)$$

In the problem where $N = 10^6$ for the conventional method, if the block size is chosen so that $(w_i + 1) = 3$, $i = 1, 2, 3$, and $S = 5$, then $N'' = 111$. If $(w_i + 1) = 4$, $i = 1, 2, 3$ and $S = 10$, a rather large set of values, then $N'' = 424$, still far less than 10^6 locations.

F. SUMMARY OF PROCEDURE

The entire state-time space over which optimal control is to be computed is partitioned into blocks according to Eq. (11). Within each time interval the blocks are ordered so that most interblock transitions will take place in the preferred direction of motion. The blocks at $t_f - \Delta T \leq t - t_0 \leq t_f$ are processed first. The points from the initialization procedure are used to begin the computations. Computation for the blocks in the time interval $(m - 1)\Delta t \leq t - t_0 \leq m\,\Delta T$ are then made on the basis of data obtained at $t = m\,\Delta T$ by the computations for blocks in the interval $m\,\Delta T \leq t - t_0 \leq (m + 1)\,\Delta T$. Within each block one of the procedures described in Section C is used. The processing continues in this manner until time t_0 is reached.

V EXAMPLES

A. MINIMUM FUEL TRAJECTORIES FOR SUPERSONIC AIR TRANSPORTS

A computer program has been written for using the new technique to compute minimum fuel trajectories for Mach 3 aircraft.(4,5,7,8) This problem is of particular interest because it is estimated that the cost of fuel may be as much as 70 percent of the total cost of a flight.(7) By inserting data for a particular airframe and engine into the program, it is possible to calculate minimum fuel trajectories for any airframe-engine configuration.

The equations of motion for the aircraft are based on quasi steady-state assumptions, as is the case in similar studies.(3, 9)

$$\frac{dv}{dt} = \frac{g}{W}(T - D) - g \sin \gamma \tag{21}$$

$$\frac{dh}{dt} = v \sin \gamma \tag{22}$$

$$\frac{dr}{dt} = v \cos \gamma \tag{23}$$

where v = velocity in direction of flight, in Mach (M)
h = altitude, in feet (ft)
r = range, in nautical miles (nm)
W = weight of aircraft plus fuel
T = thrust
D = drag
γ = flight path angle with respect to horizontal

By considering range to be the independent variable and by using Eq. (32) to relate changes in r to changes in t, it is possible to reduce the number of system equations in two. Altitude and velocity are taken to be the state variables, while thrust and flight path angle are taken to be the control variables. Drag is expressed as a function of these variables and W by the data for the specific airframe. The state transformation equations are then as follows:

$$\dot{\underline{x}} = \begin{bmatrix} \frac{dv}{dt} \\ \frac{dh}{dt} \end{bmatrix} = \begin{bmatrix} \frac{g}{W}[T - D(h, v, \gamma, W)] - g \sin\gamma \\ v \sin \gamma \end{bmatrix}$$

$$= \underline{f}(\underline{x}, \underline{u}, t) \tag{24}$$

where

$$\underline{x} = \begin{bmatrix} v \\ h \end{bmatrix} \quad \text{and} \quad \underline{u} = \begin{bmatrix} T \\ \gamma \end{bmatrix}$$

The performance criterion can be expressed as the integral of fuel flow, W_F.

$$J = \int_{t_0}^{t_f} W_F(v, h, t, \gamma, W, \sigma)\, d\sigma \tag{25}$$

$$\ell(\underline{x}, \underline{u}, t) = \frac{dW}{dt} = W_F\,(v, h, T, \gamma, W, t) \tag{26}$$

The fuel flow function is determined by the data for the specific engine.

Several constraints are present in the problem. Maximum and minimum values of thrust and flight path angle are specified by data for the airframe and engine.

$$\begin{aligned} \alpha_1^- &\leq T \leq \alpha_1^+ \\ \alpha_2^- &\leq \gamma \leq \alpha_2^+ \end{aligned} \tag{27}$$

Altitude and velocity are restricted to lie in the flight envelope determined by the airframe-engine configuration.

$$\begin{aligned} \beta_2^- &\leq h \leq \beta_2^+ \\ \beta_1^-\,(h) &\leq v \leq \beta_1^+\,(h) \end{aligned} \tag{28}$$

Other constraints that can be considered in the program include maximum sonic boom restrictions and air traffic control constraints. Variable weather conditions and emergency conditions can also be implemented.

The problem is now expressed in a form suitable for application of the new method. Since there are two state variables and two control variables, the procedure for $q \geq n$ can be used. In this procedure the admissible controls are taken to be acceleration at constant h, deceleration at constant h, climb at constant v, dive at constant v, and cruise at constants v and h. Formally,

$$\begin{aligned} U &= \left\{\underline{u}^{(2)}, \underline{u}^{(-2)}, \underline{u}^{(1)}, \underline{u}^{(-1)}, \underline{u}^{(0)}\right\} \\ &= \left\{ \begin{pmatrix} \alpha_1^+ \\ 0 \end{pmatrix}, \begin{bmatrix} \alpha_1^- \\ 0 \end{bmatrix}, \begin{bmatrix} \alpha_1^+ \\ \sin^{-1}\left(\frac{\alpha_1^+ - D}{W}\right) \end{bmatrix}, \begin{bmatrix} \alpha_1^- \\ \sin^{-1}\left(\frac{\alpha_1^- - D}{W}\right) \end{bmatrix} \begin{bmatrix} D \\ 0 \end{bmatrix} \right\} \end{aligned} \tag{29}$$

This type of altitude-velocity scheduling has been used successfully in a number of previous studies.(3, 7, 9)

The total range of altitude and velocity is $0 \leq h \leq$

90,000 ft and $0 \leq v \leq 3.00$ M. The increment sizes are $\Delta h = 1,000$ ft and $\Delta v = 0.02$ M. The block size is $2\Delta h = 2,000$ ft by $2\Delta v = 0.04$ M by $\Delta R = 10$ nm.

The preferred direction of motion for this problem is well defined; as the aircraft consumes fuel and loses weight, it tends to fly at a higher altitude and a higher velocity. Hence, the preferred direction is toward increasing velocity and increasing altitude.

The high-speed memory requirement for a single block is 54 locations. Since this value is so low, 1100 high-speed storage locations are sufficient to store the entire row of blocks that cover the same altitudes. By careful handling of the data for optimal points on the boundaries of blocks, it is possible to keep in the high-speed memory at all times all of the optimal points required for the processing of a row of blocks except for the initial set of points at $t = m\ \Delta T$. Consequently, about 3,000 optimal points are generated between referrals to low-speed memory, and less than 300 points are taken from the low-speed memory each time.

The total computing time is such that optimal trajectories over a 1,000-nm range can be generated in one hour of computing time on an existing computer.* This rate is practical for a detailed evaluation of an aircraft design or for preflight computations.

B. REAL-TIME CONTROL OF SUPERSONIC AIR TRANSPORTS

A method has been developed(4,5) for using the same basic program for real-time control of these aircraft. Although the present program provides the optimal control at all possible values of v, h, and r, it is not always sufficient to simply use a table look-up routine based on preflight computations. This situation occurs because it is possible that the conditions under which the preflight computations were made may change considerably during the flight due to changes in weather conditions, air traffic control constraints, or aircraft parameters.

---

* IBM 7090

The new program uses the results of preflight computations until it is informed that conditions have changed enough to justify a new computation. Then, using a relatively simple subroutine, it chooses a space in which it is extremely likely that the optimal trajectory will fall. This space can be regarded as the neighborhood about a nominal optimal trajectory. Computations are then performed using the basic program with the new conditions only in the blocks in this new area. If the area chosen covers 10,000 feet in altitude and 0.30 M in velocity, then the computing rate is 1000 nm. of optimal trajectory per minute.* Preliminary indications are that with this size area a satisfactory program can be written.

C. BOUNDED CONTROL OF LINEAR PLANTS WITH ARBITRARY PERFORMANCE CRITERIA

The new technique has also been applied to computing the optimal bounded control of the $1/s^2$ plant with an arbitrary performance criterion. The basic algorithm can easily be extended to other linear plants. The system differential equations are

$$\underline{\dot{x}} = \begin{bmatrix} \frac{dx_1}{dt} \\ \\ \frac{dx_2}{dt} \end{bmatrix} = \begin{bmatrix} x_2 \\ \\ u \end{bmatrix} = \underline{f}\,(\underline{x},\, u) \qquad (30)$$

The control variable is bounded

$$\alpha^- \leq u \leq \alpha^+ \qquad (31)$$

The performance criterion is taken to be the integral over $0 \leq t < \infty$ of an arbitrary time-invariant scalar function of $\underline{x}$ and u.

$$J = \int_0^\infty \ell(\underline{x},\, u)\; dt \qquad (32)$$

---

* IBM 7090

Following the method developed by Deley$^{(10)}$ for the case where $\mathcal{L}(\underline{x}, u)$ is a quadratic function of $\underline{x}$ and u, the principle of optimality is written in a form such that the optimal control is independent of the time at which control is applied.

$$I(\underline{x}) = \min_{u \in U} \left\{ \mathcal{L}(\underline{x}, u)\, \delta t + I\left(\underline{x} + \underline{f}(\underline{x}, u)\, \delta t\right) \right\} \quad (33)$$

From the equations of motion, it is clear that the optimal trajectories for almost any cost function will be clockwise motions about the origin. In order to utilize this information about the preferred direction of motion, the blocks are processed in counterclockwise rings about the origin, beginning with an initial block centered at the origin and working outward. The block size is such that nine high-speed storage locations are required per block. However, it is possible to avoid all referral to low-speed memory by storing the outermost points after computation of each ring of blocks. For a problem where 2500 locations would be necessary to store one optimal point for each possible state, 200 locations suffice to store the outermost points after computation of the largest ring about the origin.

VI CONCLUSIONS

A. COMPARISON WITH CONVENTIONAL DYNAMIC PROGRAMMING

The most important advantage of the new procedure over conventional dynamic programming is the reduction in high-speed memory requirement. This advantage is especially evident when the accuracy requirements of a problem are such that 25 or more quantization levels are required in each state variable. It has been shown that in an n-dimensional problem the high-speed memory requirement for the conventional method is of the order of N^n, where N, the number of quantization levels per state, is generally between 10 and 150; while for a new method, the requirement is of the order of w^n where w is from 2 to 5.

Other procedures have been suggested for reducing the high-speed memory requirement of conventional dynamic programming. The most widely discussed proposal is Bellman's polynomial approximation of the minimum cost function.(1, 3) This technique is effective only in those relatively well behaved problems where the minimum cost function can be approximated accurately by a low-order polynomial in the state variables. For more complicated problems where this approximation is not sufficiently good, incorrect results are obtained. On the other hand, if a high-order polynomial is used, the total computation time becomes excessive, both as a result of computing the coefficients of the polynomial (often on the basis of data in the low-speed access memory) and from repeatedly evaluating a high-order polynomial in the interpolation procedure. For these and other reasons, this technique has not been in widespread use.

An alternative procedure which has been considered(4) is to work in blocks, as in the new method, but to use the conventional method for performing computations within the block. However, because the conventional method applies a given control for a fixed time increment, it is not possible to predict how close the next state will be to the present state. Consequently, for a block which covers only a small area of state space, it can no longer be assumed that the next state is in the same block as the present state. Even if the block covers a large area of state space, at points near the boundary the same problem occurs. Techniques for introducing additional data for points in neighboring blocks require excessive referral to low-speed memory. Extrapolation of the minimum cost function is not practical because the distance in state space over which the extrapolation is made is not necessarily small. Finally, if the fixed time increment is taken to be so small that the next state is always sufficiently close to the present state, excessive computing time is required; in one example where this approach was attempted, the computing time was estimated to be 35 times that for the new procedure. As a result of these considerations, this method also appears to be of limited usefulness.

In order to compare the total computing time of the new procedure and the conventional procedure, it is necessary to consider both the time spent in actual optimization calculations and the time spent in referral to low-speed memory. For a given problem with the same set of admissible controls and the same increment sites, the two procedures take almost exactly the same amount of time for optimization calculations. If the procedure with $q \geq n$ is used, a significant saving in this time can often be realized.

If the high-speed memory requirement of conventional dynamic programming exceeds the available number of storage locations, then the number of referrals to low-speed memory is a large number for each time increment. On the other hand, the new procedure requires data from low-speed memory only once for each block, which covers several time increments. If the procedures suggested for storing the data for several blocks in the high-speed memory are used, the number of referrals is reduced further. In one previous application of the new method,(7) the time spent in referral to low-speed memory was a factor of 25 less than for the conventional method. An n-dimensional interpolation is required in both the conventional method and the general case of the new method. This interpolation is required both in computation of optimal control and in use of the results to trace out optimal trajectories. In the new method where $q \geq n$, one-dimensional interpolation (in time) is required during computation, and no interpolation is required in using the results.

The initialization procedure of the new method provides more flexibility in handling terminal constraints than does the conventional method.

The accuracy of the results obtained by the new method is the same as that for the conventional method when increment sizes, the set of admissible controls, and the accuracy of interpolation formulas are evident for both cases.

The new method is slightly more complex to program than the conventional procedure, but for the general case, the increase in complexity is no larger than 10 percent of the total effort.

The new technique, unlike the conventional method, can effectively use information about the behavior of the system to decrease both computing time and the high-speed memory requirement. As more is learned about the preferred direction of motion, the amount of computations for motion in the nonpreferred direction can be decreased. Also, if an accurate approximation to the optimal trajectory of interest can be defined, then only those blocks near this trajectory need be processed. For the same reasons that the conventional method cannot be used in blocks, it cannot be used effectively in a small region about a nominal optimal trajectory.

The new technique is applicable to most problems where conventional dynamic programming can be used. In particular, it can be applied to any problem which can be represented by a set of nonlinear time-varying differential equations, as in Eq. (1). It retains the constraint-handling capability of the conventional method, and it also computes optimal control for all possible initial states.

B. COMPARISON WITH OTHER OPTIMIZATION TECHNIQUES

The most important advantage of the new technique over both analytical techniques, such as the calculus of variations(6) and Pontryagin's maximum principle(11) and other computational techniques such as the gradient method(12) is that, like conventional dynamic programming, it is applicable to such a large class of problems. A wide variety of system equations and performance criteria can be handled within the framework of the method.

Multiple constraints of a very general nature can also be handled without difficulty.

By using information about the behavior of the system to reduce the number of computations required, the new technique can compete at a reasonable level with more specialized techniques in both computing time and high-speed memory requirement, even in relatively simple problems. In the supersonic aircraft real-time control problem, by using the new method in a region about a nominal optimal trajectory, a high-speed storage requirement, including locations for the program, was obtained which was about the same as that for the gradient method; while computing time was estimated at a little more than five times longer than that required for a single iteration of the gradient procedure. In the $\frac{1}{s^2}$ problem, the high speed storage requirement, including locations for the program, was again about the same as for Deley's method,(10) which can be used only if $\mathcal{L}(\underline{x}, u)$ is a quadratic function of $\underline{x}$ and u. The computing time was about twice that of Deley's method.

Another important consideration is that one computation with the new method provides the optimal control for all initial states in the region computed, while for most other techniques a complete recomputation is required for each initial state. This factor is important in both detailed studies and in real-time control applications.

In general the new technique is of value in problems ranging from extremely difficult examples, where no technique other than dynamic programming is applicable, to relatively simple problems with a troublesome nonlinearity or constraint. By combining the general applicability of dynamic programming with a high-speed memory requirement and computing time competitive with other existing methods, the new technique provides a powerful and useful tool for solving optimization problems.

ACKNOWLEDGMENT

Appreciation is expressed for the guidance of Prof. Gene F. Franklin and Prof. William K. Linvill of Stanford University and to Mr. D. W. Richardson and the Guidance and Control Division of Hughes Aircraft Company where the SST program was developed.

REFERENCES

1. Bellman, R., Dynamic Programming. Princeton, N. J.: Princeton University Press, 1957.

2. Bellman, R., Adaptive Control Processes. Princeton, N. J.: Princeton University Press, 1961.

3. Bellman, R., and Dreyfus, S., Applied Dynamic Programming. Princeton, N. J.: Princeton University Press, 1962.

4. Larson, R. E., "Dynamic Programming with Continuous Independent Variable," Technical Report 6302-b, Contracts Nonr 225(24), NR 373,360, SEL-64-019, Stanford University, Stanford, California (June 1964).

5. Larson, R. E., "An Approach to Reducing the High-Speed Memory Requirement of Dynamic Programming," J. Math. Anal. and Appl., (to be published).

6. Bliss, G. A., Lectures on the Calculus of Variations. Chicago: U. of Chicago Press, 1946.

7. Richardson, D. W., et al., "Control Data Investigation for Optimization of Fuel on Supersonic Transport Vehicles," Phase I Report, Contract AF 33(657)-8822, Project 9056, Hughes Aircraft Company, Culver City, California (Jan. 1963).

8. Richardson, D. W., Hechtman, E. P., and Larson, R. E., "Control Data Investigation for Optimization of Fuel in Supersonic Transport Vehicles," Phase II Report, Contract AF 33(657)-8822, Project 9056, Hughes Aircraft Company, Culver City, California (June 1963).

9. Cartaino, S., and Dreyfus, S., "Application of Dynamic Programming to the Minimum Time-to-Climb Problem," Aero. Eng. Rev., Vol. 16, No. 6 (June 1957), pp. 74-77.

10. Deley, G. W., "Optimal Bounded Control of Linear Sampled Data Systems Using Quadratic Performance Criteria," Contract NsG-133-61, Stanford Electronics Laboratory Report SEL-63-024, Stanford University, Stanford, California, (April 1963).

11. Pontryagin, L. S., et al., The Mathematical Theory of Optimal Processes, ed. by L. W. Neustadt. New York: Interscience Publishers, 1962.

12. Bryson, A. E., Jr. and Denham, W. F., "A Steepest Ascent Method for Solving Optimal Programming Problems," J. Appl. Mech., Vol. 29, Series E, No. 2 (June 1962), pp. 247-257.

SUBOPTIMAL LINEAR FILTERING FOR CONTINUOUS DYNAMIC PROCESSES

J. S. Meditch

Aerospace Corporation

Los Angeles, California 90045

I. Introduction

Considerable progress has been made in recent years in the development of the theory of optimal filtering and prediction for both continuous and discrete linear processes.(1-7) Moreover, the theory has been applied to certain problem areas with reasonable success.(3,8-11)

However, in some applications, for example, in a problem involving a large number of state variables, one may be willing to accept suboptimal performance in return for a system configuration which is less complex than the optimal one. Indeed, engineering considerations such as power, weight, volume, reliability, and/or cost may very well dictate a suboptimal design.

A theory of suboptimal linear filtering for continuous processes is developed in this paper. The present work is an extension and generalization of some results for discrete systems as presented in a recent note.(12) The method consists in partitioning the process state vector into two or more parts, and designing separate filters for each part. The central result of this paper is the derivation of a linear, first-order, matrix differential equation whose solution yields the covariance matrix of the error in the suboptimal estimate. This result provides a means of comparing suboptimal with optimal performance.

II. Optimal Linear Filtering[(2)]

The optimal linear filtering problem consists in obtaining a linear least squares estimate of a state vector x based on observations z where

$$\dot{x} = F(t)x + G(t)\, u(t) \qquad (1)$$

and

$$z(t) = H(t)x(t) + v(t). \qquad (2)$$

In (1) and (2), x is an n-vector, u is an m-vector, z and v are r-vectors, F, G, and H are continuous n x n, n x m, and r x n matrices, respectively; t denotes time and the dot denotes the time derivative. In addition, u(t) and v(t) are independent gaussian white noise processes with identically zero means and covariance matrices

$$\mathrm{cov}\,[u(t),\, u(\tau)] = Q(t) \cdot \delta\,(t - \tau)$$

$$\mathrm{cov}\,[v(t),\, v(\tau)] = R(t) \cdot \delta\,(t - \tau)$$

$$\mathrm{cov}\,[u(t),\, v(\tau)] = 0$$

for all t, τ, where δ is the Dirac delta function, Q(t) is a symmetric m x m nonnegative definite matrix, R(t) is a symmetric r x r positive definite matrix, and both matrices are assumed to be continuously differentiable in t.

It is assumed that the measurement z(t) starts at some time

t_o and that $E(x(t_o)) = 0$ where E denotes the expected value. It is further assumed that cov $[x(t_o), x(t_o)]$ is known.

The optimal estimate, denoted by $\hat{x}$, of the state x at time $t \geq t_o$, given observations up to time t, is generated by the linear dynamical system

$$\dot{\hat{x}}(t) = F(t)\,\hat{x}(t) + K(t)\,[z(t) - H(t)\,\hat{x}(t)] \tag{3}$$

with

$$K(t) = P(t)\,H^T(t)\,R^{-1}(t) \tag{4}$$

and

$$\dot{P} = F(t)P + PF^T(t) - K(t)\,H(t)P + G(t)Q(t)G^T(t) \tag{5}$$

where T denotes matrix transpose and $[\ ]^{-1}$ denotes the matrix inverse. Equation (3) is the canonical form of the optimal filter, equation (4) gives the optimal filter gain, and equation (5) is the relation for the covariance matrix P of the error in the estimate. The initial conditions on equations (3) and (5) are $\hat{x}(t_o) = E\,(x(t_o)) = 0$, and $P(t_o) = \text{cov}\,[x(t_o), x(t_o)]$, respectively.

III. Suboptimal Linear Filtering

The state vector x of the process defined by equation (1) can be partitioned into two or more parts by considering a set of constant matrices A_i, $i = 1, \ldots, M \geq 2$, and defining a corresponding set of vectors

$$y_i(t) = A_i x(t) \qquad i = 1, \ldots, M \tag{6}$$

where each A_i must have n columns. A requirement on the A_i is that

$$\sum_{i=1}^{M} A_i^{\dagger} A_i = I \tag{7}$$

where $A_i^{\dagger}$ is the pseudoinverse matrix$^{(13, 14)}$ of A_i and I is the identity matrix. (A brief discussion of the pseudoinverse matrix is given in the appendix where it is also shown that condition (7) makes the pseudoinverses unique.)

Premultiplying both sides of equation (6) by $A_i^{\dagger}$, summing over all i, and utilizing (7), one sees that

$$x(t) = \sum_{i=1}^{M} A_i^{\dagger} y_i(t). \tag{8}$$

Hence, equation (8) provides a relation for reconstructing the process state vector from the partitioned one.

Premultiplying by A_i in (1) and utilizing (8), one obtains

$$\dot{y}_i = A_i F(t) \sum_{j=1}^{M} A_j^{\dagger} y_j + A_i G(t)\, u(t), \qquad i = 1, \ldots, M \tag{9}$$

as the set of relations for the partitioned process.

It is assumed here that the measurements which served as the input to the optimal filter are also available as the input to the suboptimal filter.

For each y_i, $i = 1, \ldots, M$, the class of suboptimal filters which will be considered here are defined (by analogy to equations (3) - (5)) by the relations

$$\dot{\hat{y}}_i(t) = A_i F(t) \sum_{j=1}^{M} A_j^{\dagger} \hat{y}_j(t) + K_i(t)[z(t) - H(t) \sum_{j=1}^{M} A_j^{\dagger} \hat{y}_j(t)] \tag{10}$$

with

$$K_i(t) = P_i(t)\,[H(t)A_i^\dagger]^T\ R^{-1}(t) \tag{11}$$

and

$$\dot{P}_i = A_iF(t)A_i^\dagger\,P_i + P_i\,[A_iF(t)A_i^\dagger]^T - K_i(t)\,[H(t)\,A_i^\dagger]\,P_i$$

$$+ A_iG(t)Q(t)G^T(t)A_i^T \quad . \tag{12}$$

Equation (10) defines the canonical form of each of the suboptimal filters and equations (11) and (12) define each of the filter gains.

Each of the filters involves a suboptimal estimate $\hat{\xi}$ of x which is defined by the relation

$$\hat{\xi}(t) = \sum_{j=1}^{M} A_j^\dagger\,\hat{y}_j(t).$$

The error in the suboptimal estimate of x, which will be denoted by $\tilde{y}$, is given by

$$\tilde{y}(t) = x(t) - \sum_{i=1}^{M} A_i^\dagger\,\hat{y}_i(t) \tag{13}$$

and its derivative by

$$\dot{\tilde{y}}(t) = \dot{x}(t) - \sum_{i=1}^{M} A_i^\dagger\,\dot{\hat{y}}_i(t)\ . \tag{14}$$

Substituting into (14) from (1), (2), and (10), one obtains

$$\dot{\tilde{y}}(t) = F(t)\,x(t) + G(t)\,u(t) - \sum_{i=1}^{M} A_i^\dagger \left\{ A_iF(t) \sum_{j=1}^{M} A_j^\dagger\hat{y}_j(t) \right.$$

$$\left. + K_i(t)[H(t)\,x\,(t) + v(t) - H(t) \sum_{j=1}^{M} A_j\hat{y}_j(t)] \right\} \ . \tag{15}$$

Utilizing (7) and regrouping terms, one can easily put equation (15) into the form

$$\dot{\tilde{y}}(t) = [F(t) - \sum_{i=1}^{M} A_i^{\dagger} K_i(t) H(t)] \cdot [x(t) - \sum_{i=1}^{M} A_i^{\dagger} \hat{y}_i(t)] - \sum_{i=1}^{M} A_i^{\dagger} K_i(t) v(t) + G(t) u(t). \tag{16}$$

Now let

$$K^*(t) = \sum_{i=1}^{M} A_i^{\dagger} K_i(t) \tag{17}$$

and

$$B(t) = F(t) - K^*(t) H(t). \tag{18}$$

It then follows from (13), (17), and (18) that equation (16) assumes the form

$$\dot{\tilde{y}}(t) = B(t) \tilde{y}(t) - K^*(t) v(t) + G(t) u(t). \tag{19}$$

It is clear that $\tilde{y}$ is an n-vector.

The covariance equation for the estimation error $\tilde{y}$ will now be derived. Let $\Phi(t, \tau)$ be the transition matrix for the system of equation (19). The general solution of (19) is then given by

$$\tilde{y}(t) = \Phi(t,t_o) \tilde{y}(t_o) + \int_{t_o}^{t} \Phi(t,\tau)[G(\tau) u(\tau) - K^*(\tau) v(\tau)]d\tau \tag{20}$$

where t_o is the initial time and $t \geq t_o$. Now let $S(t)$ denote the covariance matrix of $\tilde{y}(t)$ which is given by

$$S(t) = E[\tilde{y}(t) \tilde{y}^T(t)]. \tag{21}$$

It then follows from (20) and (21) that

$$S(t) - \Phi(t,t_o) S(t_o) \Phi^T(t,t_o) =$$

$$E\left[\int_{t_o}^{t} \Phi(t,\tau)[G(\tau)u(\tau) - K^*(\tau)v(\tau)]\, d\tau \cdot \int_{t_o}^{t} [u^T(\sigma)G^T(\sigma) - v^T(\sigma) K^{*T}(\sigma)] \Phi^T(t, \sigma)d\sigma\right]$$

where $S(t_o) = \text{cov}\,[\tilde{x}(t_o), \tilde{x}(t_o)]$. Since u and v are independent gaussian white noise processes with covariance matrices $Q(t)\cdot\delta(t - \tau)$ and $R(t)\cdot\delta(t - \tau)$, respectively, the above equation simplifies to

$$S(t) - \Phi(t,t_o)\,S(t_o)\,\Phi^T(t,t_o) = \int_{t_o}^{t} \Phi(t, \tau)[K^*(\tau)R(\tau)K^{*T}(\tau) + G(\tau)$$

$$Q(\tau)G^T(\tau)]\Phi^T(t,\tau)d\tau.$$

Differentiating with respect to t, and utilizing the fact that $\Phi(t,t) = \Phi^T(t,t) = I$ for all t, one has that

$$\dot{S} - [\dot{\Phi}(t,t_o)S(t_o)\,\Phi^T(t,t_o) + \Phi(t,t_o)S(t_o)\,\dot{\Phi}^T(t,t_o)] =$$

$$K^*(t)\,R(t)\,K^{*T}(t) + G(t)\,Q(t)\,G^T(t). \qquad (22)$$

From the fact that Φ is a solution of the linear homogenous differential equation

$$\dot{\Phi} = B(t)\,\Phi$$

with $\Phi(t_o,t_o) = I$ for all t_o, it can be shown that the term in brackets on the left hand side of equation (22) is equivalent to

$$B(t)\,S + SB^T(t).$$

Hence, equation (22) becomes

$$\dot{S} = B(t)\,S + SB^T(t) + K^*(t)R(t)K^{*T}(t) + G(t)\,Q(t)G^T(t) \qquad (23)$$

with the initial condition $S(t_o)$. Equation (23) is the central result of this paper; its solution gives the covariance matrix of the error in the suboptimal estimate as a function of the original process dynamics and the suboptimal filter gains.

As remarked earlier, a suboptimal estimate of x may be defined by the relation

$$\hat{\xi}(t) = \sum_{i=1}^{M} A_i^{\dagger} y_i(t). \qquad (24)$$

Now, premultiplying in (10) by $A_i^{\dagger}$, summing over i, and using the definitions in (17) and (24), one obtains the canonical form of the complete suboptimal filter

$$\dot{\hat{\xi}}(t) = F(t)\ \hat{\xi}(t) + K^{*}(t)[z(t) - H(t)\ \hat{\xi}(t)]. \tag{25}$$

From (6), (7), (24), and the statement of the optimal linear filtering problem, it is clear that

$$\hat{\xi}(t_o) = \hat{x}(t_o) = E\ (x(t_o)) = 0. \tag{26}$$

From the definitions of S and $\tilde{y}$,

$$S(t_o) = \text{cov}\ [\tilde{y}(t_o),\ \tilde{y}(t_o)]$$
$$= \text{cov}\ [x(t_o) - \hat{\xi}(t_o),\ x(t_o) - \hat{\xi}(t_o)].$$

Since $\hat{\xi}(t_o) = 0$,

$$S(t_o) = P(t_o) \tag{27}$$

gives the initial condition on (23).

The choice of the initial conditions $P_i\ (t_o)$, $i = 1, \ldots, M$, for the set of equations in (12) is somewhat arbitrary, and is a problem deserving of further attention. For the present, it is assumed that the choice of the $P_i\ (t_o)$ is at the designer's discretion.

IV. Example

The process selected for this example is a normalized approximate model for the pitch plane dynamics of an orbiting satellite where the satellite pitch is subject to a gaussian white noise disturbance of zero mean and known variance. It is assumed that the pitch attitude, but not the attitude rate, is sensed in the presence of a second zero mean gaussian white noise of known variance which is independent of the external white noise acting on the satellite. The process equation is

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and u (t) is a zero mean scalar gaussian white noise with a constant variance which shall be taken equal to 2 for this example. The measurement is assumed given by

$$z(t) = [1\ 0]\, x(t) + v(t)$$

where z is a scalar and v(t) is a zero mean scalar gaussian white noise with a constant variance of $\frac{1}{4}$. It is assumed here that

$$P(0) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

where $t_o = 0$ is chosen as the initial time.

In terms of the notation used throughout this paper,

$$F(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad G(t) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$H(t) = [1\ 0]$, $Q(t) = 2$, $R(t) = \frac{1}{4}$, $\hat{x}(0) = 0$, and $\operatorname{cov}(x(0), x(0)) = P(0)$.

The partitioning selected is

$A_1 = [1\ -1]$ and $A_2 = [1\ \ 1]$

which, from equation (6), yields

$y_1(t) = x_1(t) - x_2(t)$ and $y_2(t) = x_1(t) + x_2(t)$.

The corresponding pseudoinverses are

$$A_1^\dagger = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} \quad \text{and} \quad A_2^\dagger = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix},$$

respectively.

For this example, equations (11) and (12) are

$$\dot{p}_1 = -p_1 - \tfrac{1}{2} k_1 p_1 + 2$$

where

$$k_1 = 2 p_1$$

and

$$\dot{p}_2 = p_2 - \tfrac{1}{2} k_2 p_2 + 2$$

where

$$k_2 = 2 p_2.$$

In these expressions, the p's and k's are scalars. The initial conditions chosen are $p_1(0) = p_2(0) = 1$.

It is easy to show that

$$p_1(t) = 1 \qquad p_2(t) = \frac{1}{2} + \frac{3}{2} \tanh\left(\frac{3}{2} t + .347\right)$$

$$k_1(t) = 2 \qquad k_2(t) = 1 + 3 \tanh\left(\frac{3}{2} t + .347\right)$$

for $t \geq 0$. Note that as $t \to \infty$, $p_2(t) \to 2$ and $k_2(t) \to 4$.

From (17),

$$K^*(t) = \frac{3}{2} \begin{bmatrix} 1 + \tanh\left(\frac{3}{2} t + .347\right) \\ -\frac{1}{3} + \tanh\left(\frac{3}{2} t + .347\right) \end{bmatrix}.$$

It then follows directly from equation (25) that the canonical form of the suboptimal filter is

$$\begin{bmatrix} \dot{\hat{\xi}}_1 \\ \dot{\hat{\xi}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\xi}_2 \\ 0 \end{bmatrix} + \frac{3}{2} \begin{bmatrix} 1 + \tanh\left(\frac{3}{2} t + .347\right) \\ -\frac{1}{3} + \tanh\left(\frac{3}{2} t + .347\right) \end{bmatrix} (z - \hat{\xi}_1)$$

with initial conditions (26) given by $\hat{\xi}_1(0) = \hat{\xi}_2(0) = 0$. For large values of t, this simplifies to

$$\dot{\hat{\xi}}_1 = \hat{\xi}_2 + 3\,(z - \hat{\xi}_1)$$

and

$$\dot{\hat{\xi}}_2 = (z - \hat{\xi}_1).$$

The covariance matrix S is 2 x 2 in this example and will be denoted by

$$S = \begin{bmatrix} s_1 & s_2 \\ s_2 & s_3 \end{bmatrix} .$$

Making the appropriate substitutions into equation (23), one obtains

$$\begin{aligned} \dot{s}_1 &= -\,(k_2 + 2)\,s_1 - 2s_2 + \frac{1}{16}\,(k_2 + 2)^2 \\ \dot{s}_2 &= -\,\frac{1}{2}\,(k_2 - 2)\,s_1 - \frac{1}{2}\,(k_2 + 2)\,s_2 + s_3 + \frac{1}{16}\,(k_2^2 - 4) \qquad (28) \\ \dot{s}_3 &= -\,(k_2 - 2)\,s_2 + \frac{1}{16}\,(k_2 - 2)^2 + 2 \end{aligned}$$

where $k_2(t) = 1 + 3\tanh\left(\frac{3}{2}t + .347\right)$ with initial conditions $s_1(0) = \frac{1}{2}$, $s_2(0) = 0$, and $s_3(0) = \frac{1}{2}$ as obtained from equation (23) and the given P(0) for this example.

The system (28) can be shown to be asymptotically stable with the equilibrium state given by

$$\bar{s}_1 = .75,\ \bar{s}_2 = 1.125, \text{ and } \bar{s}_3 = 3.375.$$

Here, $\bar{s}_1$ and $\bar{s}_3$ are the steady - state variances of the errors in the estimates $\hat{\xi}_1$ and $\hat{\xi}_2$, respectively. The standard deviations for these two estimates are then

$$\sigma_{\hat{\xi}_1} = 0.866 \qquad \text{and } \sigma_{\hat{\xi}_2} = 1.836.$$

The corresponding standard deviations of the errors in the optimal estimates can be shown to be

$$\sigma_{\hat{x}_1} = 0.771 \quad \text{and} \quad \sigma_{\hat{x}_2} = 1.295.$$

V. Discussion of Results

The example presented above was chosen primarily to illustrate the technique for suboptimal filter design and analysis. However, even in this case, it is clear that the suboptimal formulation yields a simpler implementation than the optimal formulation. For example, generation of the gains for the optimal filter for the problem given in the preceding section requires solution of the system of equations

$$\dot{p}_{11} = 2\, p_{12} - 4\, p_{11}^2$$

$$\dot{p}_{12} = p_{22} - 4\, p_{11}\, p_{12}$$

$$\dot{p}_{22} = -\, 4\, p_{11}^2 + 2$$

with the gains given by $k_1(t) = 4\, p_{11}\,(t)$ and $k_2\,(t) = 4\, p_{12}(t)$. In contrast to this, the suboptimal formulation requires solution of the two "uncoupled" equations

$$\dot{p}_1 = -\, p_1 - p_1^2 + 2$$

and

$$\dot{p}_2 = p_2 - p_2^2 + 2$$

to generate the gains. Thus, for this example, it is clear that the suboptimal approach reduces a two-dimensional problem to two one-dimensional problems.

Computationally then, a very attractive feature of the suboptimal filtering approach given here is that it permits one to reduce a problem of given dimension into two or more problems of lower dimension which can be solved separately.

It should be noted that the suboptimal filter formulation divorces the synthesis problem from the analysis problem whereas the two are integral in the optimal filtering formulation. This is seen by noting that optimal filter synthesis is specified by equations (3) - (5) with the covariance matrix of the estimation error being the solution of equation (5). On the other hand, the suboptimal filter synthesis is given by equations (10) - (12), but the covariance matrix of the estimation error is governed by equation (23).

It is clear that the choice of the partitioning matrices A_i is arbitrary for a given problem. For example, the choice

$$A_1 = [1 \ 0] \qquad \text{and} \qquad A_2 = [0 \ 1]$$

with

$$A_1^{\dagger} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad A_2^{\dagger} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is perfectly legitimate for the problem considered in the preceding section. Unfortunately, this particular choice leads to a filter configuration for which the variance in the error in the estimate $\hat{\xi}_2$ increases linearly with time. This partitioning is clearly undesirable since it yields a partitioned process which is not observable.(2)

In general then, it is to be expected that the choice of the A_i will require some trial and error analyses to determine which one of a number of suboptimal filters gives satisfactory performance.

Appendix

A complete treatment of pseudoinverse matrices is available in the literature.(13, 14) Only those details relevant to the present work are reviewed here.

A matrix $A^{\dagger}$ is termed a pseudoinverse of a rectangular matrix A if it satisfies the relation

$$A A^{\dagger} A = A. \tag{I}$$

If A is square with nonzero determinant, then it is clear that $A^{\dagger} = A^{-1}$. It has been established(13, 14) that, for a given A, $A^{\dagger}$ exists, but is not unique as defined by (I).

Now consider a set of constant matrices A_i, $i = 1, \ldots, M \geq 2$ and define a set of vectors y_i by

$$y_i = A_i x \qquad i = 1, \ldots, M \tag{II}$$

where x is an n-vector, y_i is an r_i-vector. ($r_i < n$; $i = 1, \ldots M$), and the A_i are $r_i \times n$ matrices.

Let $A_i^{\dagger}$ be a pseudoinverse of A_i for each i as defined in (I). In addition, assume that the $A_i^{\dagger}$ are chosen so that

$$\sum_{i=1}^{M} A_i^{\dagger} A_i = I \tag{III}$$

where I is the n x n identify matrix.

It will now be shown that the $A_i^{\dagger}$ are unique.

Assume that for each A_i there exists a second pseudoinverse $B_i^{\dagger} \neq A_i^{\dagger}$, and, further, that this set of matrices also satisfies (III), i.e.,

$$\sum_{i=1}^{M} B_i^{\dagger} A_i = I. \tag{IV}$$

It then follows immediately from (III) and (IV) that

$$\sum_{i=1}^{M} (B_i^{\dagger} - A_i^{\dagger})\, A_i = 0. \tag{V}$$

Since the choice of the A_i is arbitrary in the partitioning of x, it is clear that

$$B_i^{\dagger} = A_i^{\dagger} \qquad i = 1, \ldots, M,$$

and the proof is complete.

References

1. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, Series D, J. Basic Engr., vol. 82, 1960, pp. 35 - 45.

2. Kalman, R. E., and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. ASME, Series D, J. Basic Engr., vol. 83, 1961, pp. 95 - 108.

3. Battin, R. H., "A Statistical Optimizing Navigation Procedure for Space Flight," ARS Jour., vol. 32, 1962, pp. 1681 - 1692.

4. Ho, Y. C., "The Method of Least Squares and Optimal Filtering Theory," RM-3329-PR, The Rand Corp., Santa Monica, Calif., Oct., 1962.

5. Ho, Y. C., "A Bayesian Approach to Problems in Stochastic Estimation and Control," 1964 Joint Automatic Control Conf., Stanford Univ., Stanford, Calif., June, 1964.

6. Bryson, A. E., Jr., and D. E. Johansen, "Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise," Joint AIAA - IMS - SIAM - ONR Symposium on Control and System Optimization, U. S. Naval Postgraduate School, Monterey, Calif., Jan., 1964.

7. Cox, H., "Estimation of State Variables via Dynamic Programming," 1964 Joint Automatic Control Conf., Stanford Univ., Stanford, Calif., June, 1964.

8. Gunckel, T. L., "Orbit Determination Using Kalman's Method," Inst. of Nav. Conf., Anaheim, Calif., April, 1963.

9. Smith, G. L., S.F. Schmidt, and L. A. McGee, "Application of Statistical Filter Theory to the Optimal Estimation of Position and Velocity On Board a Circumlunar Vehicle," NASA TN D-1208, National Aeronautics and Space Administration, Ames Research Center, Mt. View, Calif., 1962.

10. Denham, W. F., and J. L. Speyer, "Optimal Measurement and Velocity Correction Programs for Midcourse Guidance," AIAA Jour., vol. 2, no. 5, May, 1964, pp. 896 - 907.

11. Kumar, K. S. P., and R. Sridhar, "On the Identification of Linear Systems," 1964 Joint Automatic Control Conf., Stanford Univ., Stanford, Calif., June 1964.

12. Joseph, P. D., "Sub-Optimal Linear Filtering," IOC 9321.4-653, Space Technology Lab., Inc., Redondo Beach, Calif., Dec. 5, 1963.

13. Penrose, R., "A Generalized Inverse for Matrices," Proc. Cambridge Phil. Soc., vol. 51, 1955, pp. 466 - 413.

14. Penrose, R., "On Best Approximate Solutions of Linear Matrix Equations," Proc. Cambridge Phil. Soc., vol. 52, 1956, pp. 17 - 19.

SYNTHESIS OF APPROXIMATELY OPTIMUM CONTROL SYSTEMS BY CONVENTIONAL BODE AMPLITUDE DIAGRAMS

A. P. Sage
Department of Electrical Engineering
University of Florida
Gainesville, Florida

R. B. Streets
The Boeing Company
Seattle, Washington

INTRODUCTION

This paper considers linear autonomous systems with stationary random inputs which are optimum in the sense that the mean square error between desired output and actual output is minimized.

The basic solution to this problem was developed some twenty years ago by Wiener. Extensions to Wiener's theory have dealt with constraints imposed by the characteristics of physical systems and multiple input output problems. Other extensions have removed one or more basic assumptions in the basic theory, such as stationary. Serious obstacles to the practical application of this theory exist because of the mathematical sophistication and algebraic complexity necessary in applying the solution technique.

This paper develops a technique which yields to a very good approximation the open loop transfer function of Wiener systems for stationary random inputs and linear autonomous fixed plants. The open loop transfer functions are constructed on Bode amplitude diagrams using the following rules.

When the normalized signal spectral density is greater than the normalized noise spectral density, the squared magnitude of the open loop transfer function looks like the normalized signal spectral density. Sufficient Butterworth zeros and poles are then added such that the open loop system had adequate phase margin and such that pole zero excess requirements as set forth by Darlington are satisfied.

STATEMENT OF THE PROBLEM

The basic assumptions of the Wiener Theory(1,2,3) are used. The control system is optimum in the sense of minimizing the mean square or integral square error. The systems are linear and time invariant. In addition the following assumptions, which are satisfied in most control problems are made. The desired output is the input signal. There is no crosscorrelation between inputs. These two assumptions are easily removed. It is also assumed that the RMS error is finite and is less than one half the RMS value of the signal. Also $\theta(W) \leq \theta(G_f) \leq -1$. ($\theta(W)$ is the zero pole excess in $H(s)$) This assumption is necessary to avoid meaningless solutions.

Spectral densities of the input signal Φ_{SS}, input noise Φ_{NN} and output disturbance Φ_{DD} are given as well as the transfer function of the fixed plant, linear region $\pm L$ of any saturating nonlinearity, and constraint equations $E\left\{i_n^2\right\} \leq J_n$. It is desired to find the optimum Gc and H to minimize $E\{e^2\}$ with all $E\left\{i_n^2\right\} \leq J_n$ unless a load disturbance is present, there is no loss in generality in setting H = 1.(4) The basic block diagram of the optimization problem is shown in Figure 1. The open loop transfer function G will represent the product $G_c G_f$ and the closed loop transfer function will be called $W = \frac{G}{1+G}$. When no confusion will result, functions of the variable (s) are written as $F(s) = F$.

APPROXIMATE WHITE NOISE OPTIMUM FILTER

It is assumed that $\Phi_{SS} = \frac{N}{D}$ and $\Phi_{NN} = 1$ (normalized) where N and D are each polynomials in s. The conventional Wiener filter is determined by the operations

$$W(s) = \frac{1}{[\Phi_{SS} + \Phi_{NN}]_+} \left\{ \frac{\Phi_{SS}}{[\Phi_{SS} + \Phi_{NN}]_-} \right\}_{PR} \qquad -1-$$

where $[\;]_+$ is the spectral factorization operator denoting the extraction of the multiplicative term containing poles and zeros in the left half plane only. $\{\;\}_{PR}$ is the physically realizable operator defined by

$$\{\Phi(s)\}_{PR} = \frac{1}{2\pi j} \int_0^{\infty} \int_{-j\infty}^{j\infty} \Phi(p) e^{pt} e^{-st} \, dp \, dt \qquad -2-$$

For this problem, the optimum transfer function becomes

$$W = \frac{D_+}{[D+N]_+} \left\{ \frac{N}{D_+ [D+N]_-} \right\}_{PR} \qquad -3-$$

$$= \frac{D_+}{[D+N]_+} \left\{ \frac{[D+N]_+ - D_+}{D_+} + \frac{[D+N]_- - D_-}{[D+N]_-} \right\}_{PR}$$

$$W = \frac{[D+N]_+ - D_+}{[D+N]_+} \qquad -4-$$

$$G = \frac{W}{1-W} = \frac{[N+D]_+ - D_+}{D_+} = \left[\frac{N}{D} + 1\right]_+ - 1 \qquad -5-$$

This shows that when the signal spectral density is greater in magnitude than the noise spectral density, the open loop transfer function is almost the same as $[\Phi_{SS}]_+ = \left[\frac{N}{D}\right]_+$ the part of the factored normalized signal spectral density which contains poles in the left and right half plane only.

Consider the case where $\Phi_{SS} = \frac{1}{(-\lambda)^{2n}}$, $\Phi_{NN} = 1$, $\lambda = \frac{s}{\omega_c}$

$$G = \left[\frac{1}{(-\lambda)^{2n}} + 1\right]_+ - 1 = \frac{[(-\lambda)^{2n} + 1]_+ - \lambda^n}{\lambda^n} \qquad -6-$$

but

$$1 + (-\lambda)^{2n} = B_n(\lambda) B_n(-\lambda) \qquad -7-$$

Where $B_n(\lambda)$ is the n^{th} order Butterworth Polynomial. It can be shown(5) that to a very good approximation

$$B_{n-1}(\sqrt{2}\,\lambda) \approx B_n(\lambda) - \lambda^n \qquad -8-$$

The solution to this example is approximately

$$G(\lambda) \approx \frac{B_{n-1}(\sqrt{2}\,\lambda)}{\lambda^n}$$

In the vicinity where the signal and noise spectral densities are approximately equal or where the noise spectral density is greater than the signal it is always possible to approximate

$$\Phi_{SS} \approx \frac{1}{(-\lambda)^{2n_K}}, \quad \Phi_{NN} = 1 \qquad -9-$$

and use for the optimum open loop transfer function in this region

$$G \approx \frac{B_{n_K-1}(\sqrt{2}\,\lambda)}{\lambda^{n_K}} \qquad -10-$$

Amplitude and frequency normalization have been used such that "cross-over" (frequency where signal and noise spectral densities are both unity) occurs at λ equals one radian per second. The approximation of the open loop transfer function of the optimum system is therefore completed by adding Butterworth zeros at $\omega = \frac{\omega_c}{\sqrt{2}}$ such that $\theta(G) = -1$. This satisfies the pole zero excess requirements for optimum systems as set forth by Darlington.(6)
Thus

$$G_{opt}(s) \cong \left[\Phi_{SS}\right]_+ B_{n-1}\left(\sqrt{2}\,\frac{s}{\omega_c}\right) \qquad -11-$$

where n is the Bode diagram slope of $\left[\Phi_{SS}\right]_+$ at crossover.

EXAMPLE OF WHITE NOISE OPTIMIZATION

It is given that $\Phi_{SS} = \frac{A^2}{\left(1-\frac{s^2}{b^2}\right)^2}$, $\Phi_{NN} = N^2$ and it is desired to determine the optimum filter. The normalized signal spectral density is $\Phi_{SS}'(s) = \frac{\left(\frac{A}{N}\right)^2}{\left(1-\frac{s^2}{b^2}\right)^2}$ which has been a crossover frequency (spectral density equals unity) of $\omega_c = b\sqrt{\frac{A}{N}}$. The optimum open loop filter is

$$G \approx \frac{\frac{A}{N}}{\left(1+\frac{s}{b}\right)^2}\left(1+\sqrt{2}\,\frac{s}{\omega_c}\right) \qquad \omega_c = b\sqrt{\frac{A}{N}}$$

The exact solution can be obtained from equation 1 or 5 and is

$$G_{opt}(s) = \frac{\left(\sqrt{1+\left(\frac{A}{N}\right)^2}-1\right)\left(1+\sqrt{2}\,\frac{s}{\omega_a}\right)}{\left(1+\frac{s}{b}\right)^2}$$

$$\omega_a = b\left(\sqrt{1+\sqrt{1+\left(\frac{A}{N}\right)^2}}+\sqrt{2}\right)$$

Figure 2 shows the steps in the construction of the Bode Amplitude Diagram for the determination of the optimum filter. Figure 3 and 4 show the error and percentage approximation error (compared to the exact optimum) in using the approximate filter. The approximation error is less than 4% if the RMS error is less than 1/2 the RMS signal. Certainly this should be true for control applications.

EXTENSION TO NON WHITE NOISE INPUTS, MINIMUM BANDWIDTH SYSTEMS, AND SATURATION CONSTRAINTS

In order to satisfy the requirement that $\theta(W) \le \theta(G_f) = -n \le -1$, consider a noise spectral density $\Phi_{NN} = (-b^2 s^2)^{n-1}$ where b is arbitrary. In order to satisfy a bandwidth constraint, $\omega_c < b$, b is determined by $\Phi_{SS}(jb) = \Phi_{NN}(jb)$.

For a signal density of the form

$$\Phi_{SS} = \frac{1}{(-\lambda)^{2K}}\,, \quad \Phi_{NN} = (-\lambda)^{2m}\,, \quad K \geq 1\,, \quad m \geq 0 \qquad \text{-12-}$$

it can easily be shown by application of equation 1 that

$$W = \frac{B'_{K+m}(\lambda)}{B_{K+m}(\lambda)}\,, \qquad G = \frac{B'_{K+m}(\lambda)}{B''_{K+m}(\lambda)} \qquad \text{-13-}$$

where $B'_{K+m}(\lambda)$ denotes the first K terms of the normalized Butterworth polynomial and $B''_{K+m}(\lambda)$ denotes the last $m+1$ terms of the $K+m^{th}$ order Butterworth polynomial. The approximation rules (with normalization such that $\omega_c = 1$ and $\Phi_{SS}(j\omega_c) = \Phi_{NN}(j\omega_c) = 1$)

for other than white noise inputs can be stated as:

a) For $\Phi_{ss} > \Phi_{nn}$

$$G \approx [\Phi_{ss}]_+$$

b) For $\Phi_{ss} \approx \Phi_{NN}$, $\text{slope}[\Phi_{ss}(j\omega_c)]_+ = -K$, $\text{slope}[\Phi_{NN}(j\omega_c)]_+ = m$

$$G \approx \frac{B'_{K+m}(\lambda)}{B''_{K+m}(\lambda)}$$

c) For $\Phi_{ss} < \Phi_{NN}$

$$\theta(G[\Phi_{NN}]_+) = -1$$

These three rules allow approximate calculation of G for a variety of problems. All single input single output optimization problems involving minimum phase linear fixed plants and quadratic error criteria can be cast into the framework of a Wiener problem with a colored noise input.

EXAMPLE 1 OF SATURATION CONSTRAINT

It is given that $\Phi_{ss} = \frac{1}{-s^2}$, $\Phi_{NN} = 0$, $G_f = \frac{1}{s^2}$, $\sigma_i^2 < J$.

It is easily shown that the saturation constraint problem can be treated as a conventional Wiener problem where the "noise" spectral density is considered as

$$\Phi_{NN}' = \Phi_{NN} + \frac{i^2[\Phi_{ss} + \Phi_{NN}]}{G_f(s)\,G_f(-s)}$$

and i is adjusted to satisfy the saturation constraint. For this particular example, the equivalent noise spectral density is

$$\Phi_{NN}' = -i^2 s^2$$

The spectral densities are amplitude normalized by a factor i and then frequency normalized, $\lambda = \sqrt{i}\,s$ to give

$$\Phi_{ss\,Norm} = \frac{1}{-\lambda^2} \quad , \quad \Phi_{NN\,Norm} = -\lambda^2$$

The optimum filter is easily determined using the colored noise rules and is

$$G_{opt}(\lambda) = \frac{1}{\lambda(\sqrt{2} + \lambda)}$$

or

$$G_{opt}(s) = \frac{1}{\sqrt{i}\,s\,(\sqrt{2} + \sqrt{i}\,s)}$$

The mean square saturation signal is calculated from the saturation signal spectral density

$$\Phi_{ii} = \frac{\Phi_{SS} + \Phi_{NN}}{G_f(s)\,G_f(-s)}\,W(s)\,W(-s)$$

which gives

$$\Phi_{ii} = \frac{-s^2}{(1 + \sqrt{2i}\,s + i s^2)(1 - \sqrt{2i}\,s + i s^2)}$$

Consequently $\sigma_i^{\,2} = \frac{1}{2\sqrt{2i}\,i} = J$

In this example the solution obtained by the approximate technique is also the correct solution.

Considerable simplification can result in the calculation of the mean square saturation signal if it is also accomplished by the Bode diagram techniques used to determine the approximately optimum filter. The equivalent noise for the saturation problem is

$$\Phi_{NN}' = \frac{i^2[\Phi_{SS} + \Phi_{NN}]}{G_f(s)\,G_f(-s)} + \Phi_{NN}$$

Therefore for $\omega > \omega_c$, $\theta(G[\Phi_{NN}']_+) = \theta(W[\Phi_{NN}']_+) = -1$, the saturation signal spectral density is

$$\Phi_{ii} = \frac{[\Phi_{SS} + \Phi_{NN}]}{G_f(s)\,G_f(-s)}\,W(s)W(-s) = \frac{[\Phi_{NN}' - \Phi_{NN}]\,W(s)W(-s)}{i^2}$$

and thus $\theta([\Phi_{ii}]_+) = -1.$ The saturation constraint consists of adding a suitable Butterworth break to Φ_{ii} at the saturation crossover frequency and adjusting the crossover frequency until the constraint is satisfied.

EXAMPLE 2 OF SATURATION CONSTRAINT

It is given that $\Phi_{ss} = \frac{625}{1-s^2}$, $\Phi_{NN} = 1$, $G_f = \frac{1}{s(1+s)}$, $J = \frac{(25)^2}{2} \times 10^8$

The first step in the solution is to determine the Lagrange multiplier, i . This is accomplished by determining the optimum approximate closed loop transfer function for no constraint $\left(W = \frac{1}{1+.04s}\right)$ and then the saturation signal spectral density for no constraint

$$[\Phi_{ii}]_+ = \frac{[\Phi_{ss} + \Phi_{NN}]_+}{G_f(s)} W(s) = 25\, s$$

This is bounded by the -1 slope as shown in the dotted line of Figure 5. Thus

$$[\Phi_{ii}]_+ = \frac{25\, s}{1 + \sqrt{2i}\, s + i\, s^2}, \qquad \sigma_i^2 = \frac{(25)^2}{2} \times 10^8 = \frac{(25)^2}{2\sqrt{2i}\, i}$$

therefore

$$\frac{1}{\sqrt{i}} = 521$$

In determining the optimum filter, the equivalent noise has two components, one of which is due to the saturation constraint and has the spectral density

$$\left[\frac{i^2 [\Phi_{ss} + \Phi_{NN}]}{G_f(s)\, G_f(-s)}\right]_+ = i\, 25\, s\, (1 + .04\, s)$$

The approximately optimum system subject to the constraint can be determined by inspection of the Bode diagram of Figure 6, and is

$$G_{Approximate} = \frac{25}{(1+s)\left(1 + \frac{\sqrt{2}\, s}{521} + \left[\frac{s}{521}\right]^2\right)}$$

(7)

The exact solution can be found in Coales, and is

$$G_{exact} = \frac{22.5}{(1+s)\left(1 + \frac{\sqrt{2}\, s}{534} + \left[\frac{s}{534}\right]^2\right)}$$

EXTENSION TO THE REGULATOR PROBLEM

This method can be applied with success to the infinite time regulator problem. The behavior of the fixed plant is described by the differential equation

$$\dot{\underline{x}} = A\underline{x} + \underline{b}\,u \qquad -14-$$

$$\underline{x}(0) = \underline{x}_0$$

where $\underline{x}$ and $\underline{b}$ are n vectors, A is constant n x n matrix and u is a scalar. The system is assumed to be completely controllable and observable. The performance index is

$$PI = \int_0^\infty (\underline{x}^T Q \underline{x} + u^2)\,dt \qquad Q = R^T R \qquad -15-$$

$\underline{k}_\infty$ is a fixed control vector such that $u = -\underline{k}_\infty^T \underline{x}$. It can be shown(8) that a necessary and sufficient condition that $\underline{k}_\infty$ be an optimal control vector is that

$$\left[1 + \underline{k}_\infty^T \Phi(s)\underline{b}\right] = \left[1 + F^T(-s)F(+s)\right]_+ \qquad -16-$$

where

$$\Phi(s) = (sI - A)^{-1} \qquad -17-$$

$$F(s) = R\,\Phi(s)\,\underline{b} \qquad -18-$$

$k_\infty^T \Phi(s)\underline{b}$ is just the loop gain of the block diagram of Figure 1, GH. Therefore

$$GH_{opt} = \left[1 + F^T(-s)F(+s)\right]_+ - 1 \qquad -19-$$

This equation is exactly the equivalent of equation 5, the White Noise optimum filter equation. Thus this technique can be used to determine the optimum feedback gains for the Kalman regulator (dual) problem

SUMMARY

This paper presents a new approach to an old solution for the stationary optimization of linear systems. In particular it represents a computational contribution to the Wiener Theory. It is shown that the infinite terminal time linear regulator problem can be solved using the white noise theory. If the Kalman gains $\underline{k(t)}$ for finite terminal time can be successfully approximated using the infinite interval gains, this method may prove advantageous for nonstationary optimization as well.

REFERENCES

(1) N. Weiner Extrapolation, Interpolation and Smoothing of Stationary Time Seves. John Wiley and Sons, New York, 1949.

(2) G. C. Newton, L.A. Gould, J.F. Kaiser Analytical Design of Linear Feedback Controls. John Wiley and Sons, New York, 1957.

(3) S.S.L. Chang, Synthesis of Optimum Control Systems, McGraw Hill Book Co., Inc. New York, 1961.

(4) J. L. Bower and P.M. Schultheiss, Introduction to the Design of Servomechanisms, John Wiley and Sons, New York, 1958.

(5) R. B. Streets, "Integration of Analytical and Conventional Design Techniques for Optimum Control Systems," Ph.D. thesis, University of Arizona, 1964.

(6) S. Darlington, "Linear Least Squares Smoothing and Prediction with Applications" Bell System Technical Journal Vol. 37 Number 5, pp. 1221-1294, 1958.

(7) J. F. Coales and P. J. Lawrence "Preparation of Charts and Tables for the Optimization of Automatic Control Systems with Random Inputs", Automatic and Remote Control, Vol. 2 Edited by J. F. Coales, Butterworths, London 1961. pp. 753-760.

(8) R. E. Kalman "When is a Linear Control System Optimal" Journal of Basic Engineering, American Society of Mechanical Engineers March, 1964.

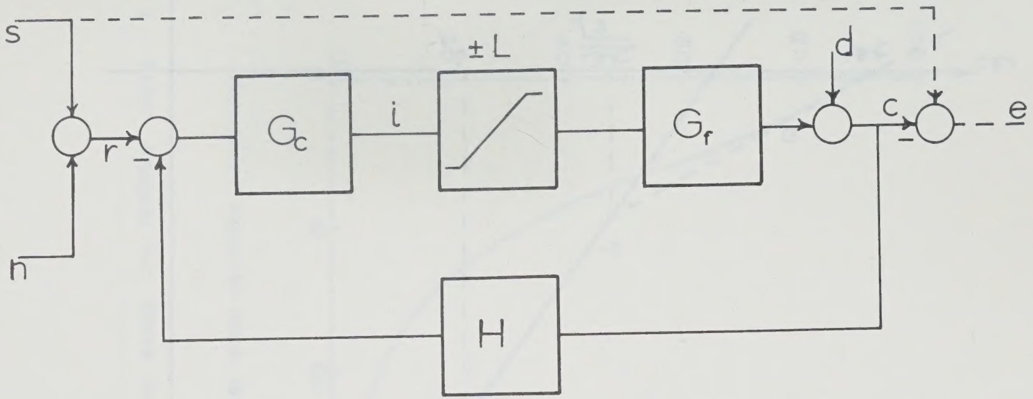

s
h
r
i
±L
d
c
e
G_c
G_r
H

Figure 1 Basic Block Diagram

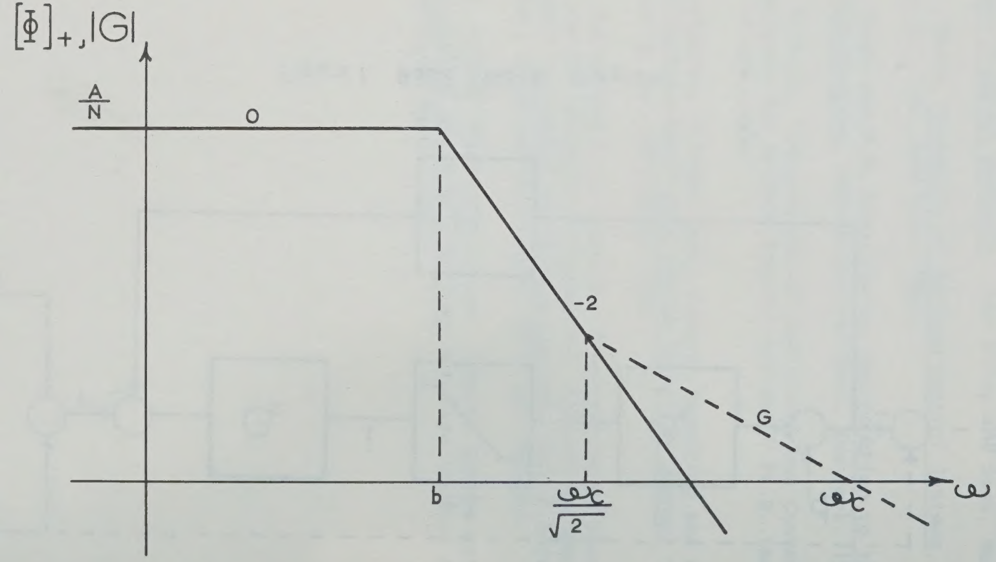


Figure 2 Bode Diagram for White Noise Example

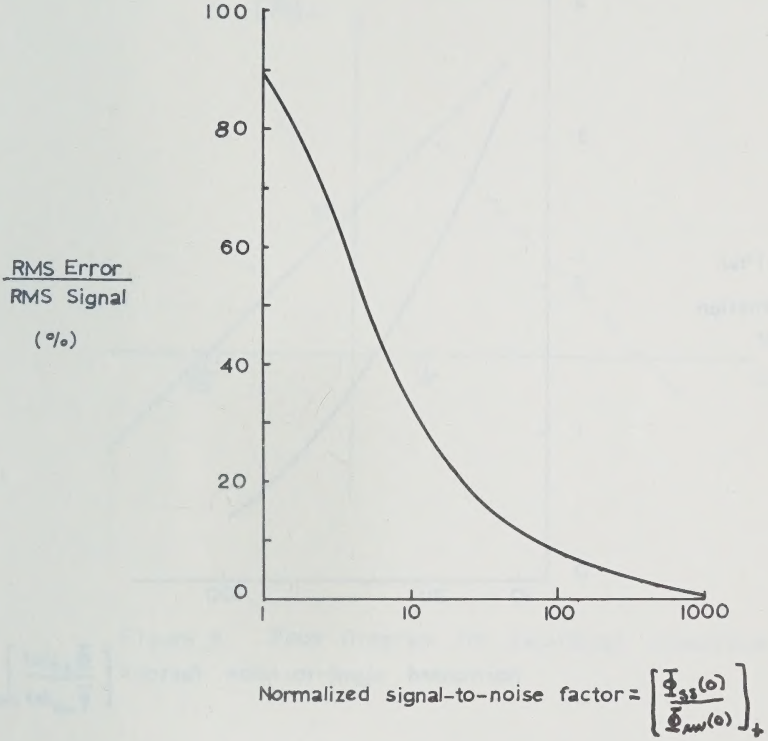


Figure 3 Error Analysis for White Noise Example

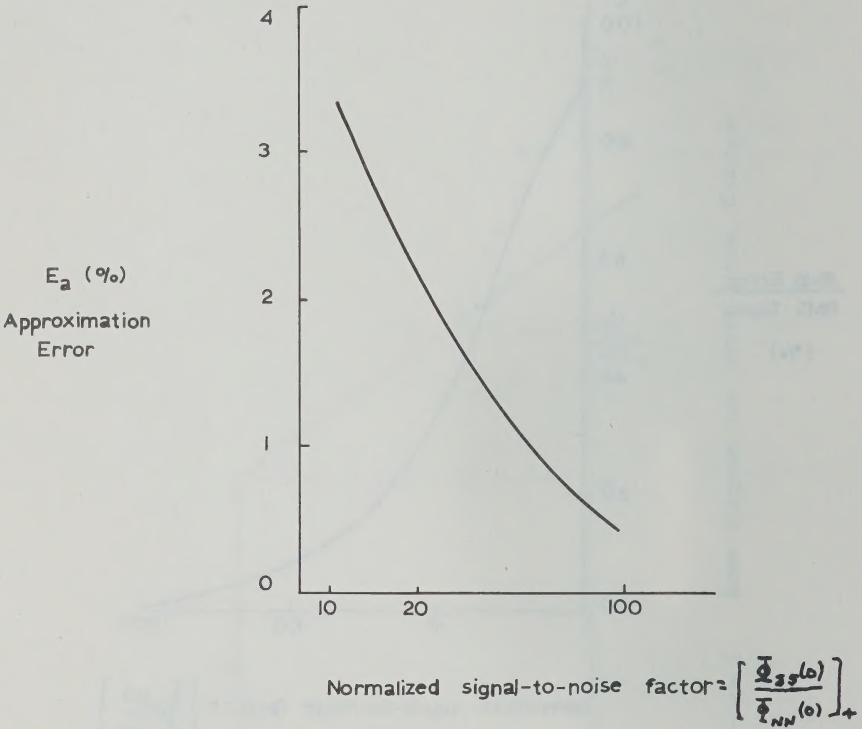


Figure 4 Approximation Error for White Noise Example

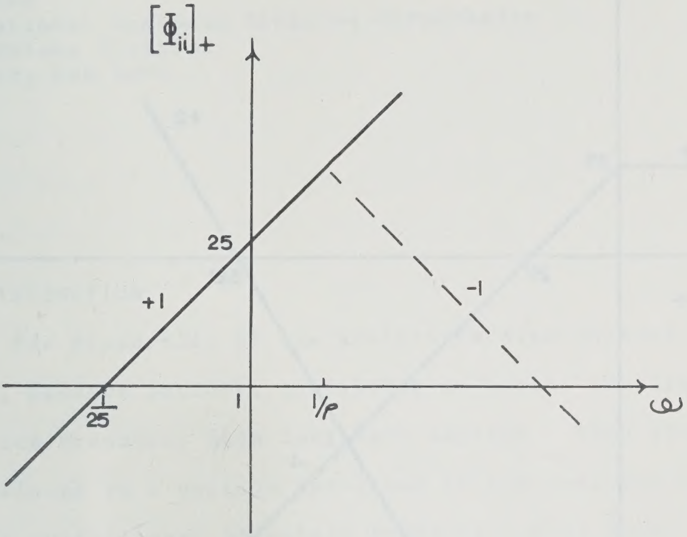


Figure 5 Bode Diagram for Saturation Constraint

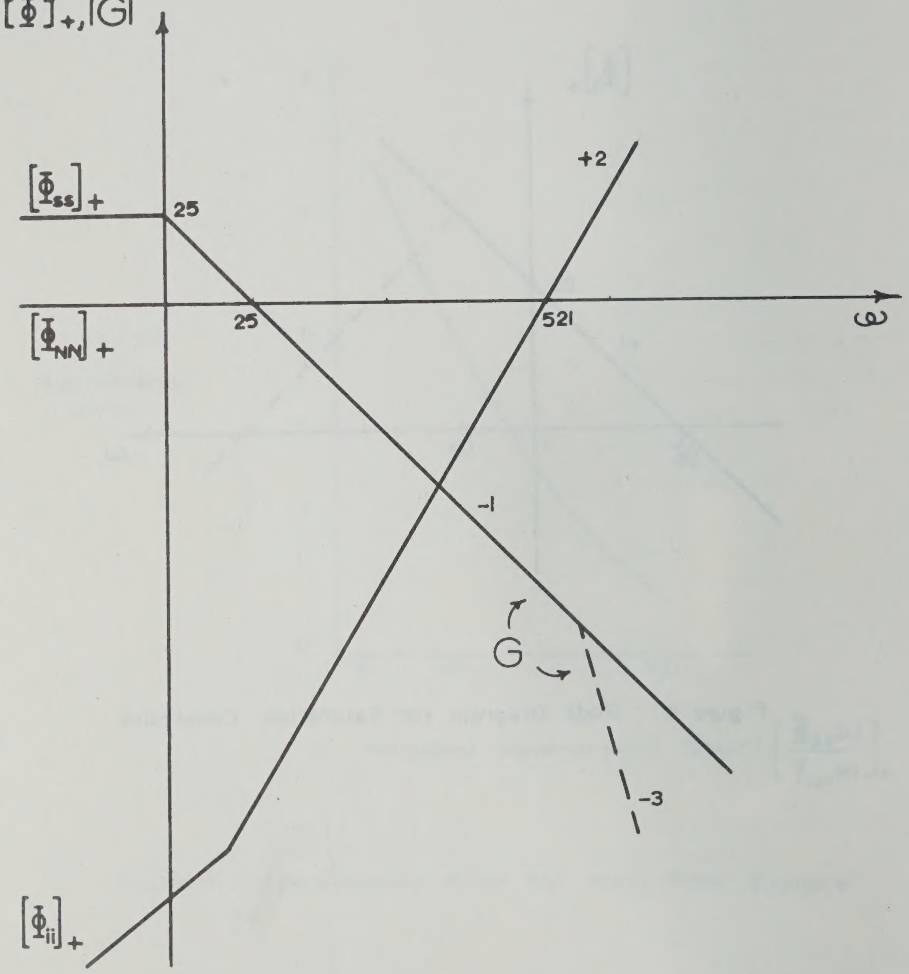


Figure 6 Bode Diagram for optimum G(s) with saturation

CERTAIN CONSTRAINTS IN THE ADMITTANCE DISCRIMINANTS OF NETWORKS*

H. T. Lee
International Business Machines Corporation
Data Systems Division
Kingston, New York

I. INTRODUCTION

A few properties of the admittance discriminant of a linear, passive network, containing no mutual coupling among its branches, have long been exposed. This discriminant belongs to a certain sub-class of the positive homogeneous multi-linear algebraic forms (p.h.m.l. form) in the network branch admittances. For a well known reason,(1) members of this sub-class of p.h.m.l. forms shall be referred to as complete tree-sum forms, or simply tree-sum forms (t.s.forms). Accordingly, any p.h.m.l. form shall be called a tree-sum-type form (t.s.t. form), before it can be determined whether it is truly a t.s. form.

---

*Except for limited amount of additions and revisions of certain procedures, as will be marked by ** in the text, the content of this paper is essentially taken from Section II and III of a dissertation, submitted in partial fulfillment of the requirement of the D.E.E. degree at the Polytechnic Institute of Brooklyn, June, 1960.

Most investigations of the past concerning the admittance discriminants had been directed toward the task of determining them or their related functions, having given the networks.(1,8) Practically no significant work had been reported on the converse problem, that of determining the realizability of a given p.h.m.l. form as an admittance discriminant of a physical network.(9) It is in the study of such a problem that the subject material of this paper was developed. A point of distinction should be made here. The object for investigation in this study is the total discriminant, and it is to be handled basically in its entirety. This problem is not of the same nature as that of many reported kindred studies, wherein switching circuits and one or another derivable entity from the discriminant were dealt with.(8,10-16) Primarily designed to serve as a set of tools for systematic decomposition and cross examination of a given t.s.t. form, this material consists largely of an exposition of certain algebraic operations and criteria which will be of use at various stages in the solution of a practical problem. In a theoretical work, it gives the rules for treating an arbitrary t.s.t. form (or its derived form) and substantiates the exerted consequences. Of particular relevance to this presentation, however, is the fact that these algebraic operations, through the topological interpretation of their graphical counterparts, do bring into light a number of constraints in the structure of realizable p.h.m.l. forms. They are exhibited as necessary conditions for realizability immediately associated with the specific operations. Besides them, there are some realizability conditions which can only be revealed through systematic application of certain elementary operations and rules, and others which are not directly related to any algebro-graphical operation but are useful means, and still others which are merely outcomes accrue to one or another conceptual artifice introduced for convenience in the original study. All these are assembled here for the dual purpose of providing a ready reference for a projected subsequent use and of promoting an interest in the detailed structural properties of a t.s. form.

II. PREPARATORY REMARKS

To save space, the entire exposition is made in a nearly itemized manner whenever possible. Most of the realizability conditions will be stated without elaboration or proof, as the majority of them are rather obvious once the points are made. It is taken for granted that the reader is familiar with the usual terms in the theory of linear graphs and their custom employment in the networks theory. Hence, despite the title given to this paper, the content will be so expressed as to comply with current practice.

To initiate the presentation, it is expedient to set forth the object for discussion in precise terms, both graphically and algebraically. Thus, consider a labeled linear graph, G, consisting of N vertices and B branches, each terminating on two distinct vertices. For such a graph, a set of distinct trees is defined and a complete sum of the tree-products formed,

$$T = \sum_{i=1}^{\nu} T^{(i)} = \sum_{i=1}^{\nu} \left[\prod_{j=1}^{n=N-1} b^{(ij)} \right]$$

where $T^{(i)}$ is the i^{th} tree-product (or simply i^{th} tree) in G, and $b^{(ij)}$ is its j^{th} branch. n is referred to as the order of T, and ν, the size of T. This T is the counterpart of the admittance discriminant in network analysis. Algebraically, the features of T can be put no further than the insufficient descriptive classification "positive homogeneous multi-linear form," where positiveness should be understood to mean that the coefficient of each team in T is + 1. Indeed, the rudimentary algebraic description falls far too short of characterizing a true t.s. form, T. For any non-trivial case, there are many other constraints in T which, when properly oriented in coexistence, distinguish T from among a host of un-realizable p.h.m.l. forms of similar size and order in the same set of variables. An un-realizable p.h.m.l. form when known to be so, shall be called a non-tree-sum form, symbolized by $\overline{T}$. Furthermore, any arbitrary p.h.m.l. form of adequate appearance shall be called a t.s.t. form, symbolized by t, before it can be determined to be a T or a $\overline{T}$. It is such t and T that are made the objects of the present study.

Before going into the subject matter, it is desirable to clarify one point as it will undoubtedly occur in the reader's mind as he proceeds. As will be seen, every considered algebraic observation or operation on a T carries with it the meaning and consequence of a specific graphical counterpart on the corresponding G. Now, when the same is applied on a t, of which the realizability is still in question, what can one say of its meaning or consequence? Undoubtedly, its meaning is primarily purely algebraic with its graphical significance contingent to the actual realizability of t. As for the consequence, as long as this algebraic encounter does not actually or tend to rationalize a $\overline{T}$ into a T, one may very well treat the t or its derived form subsequent to the pertinent algebraic encounter as if it would have a proper graphical significance and make further realizability demands on it without danger of prejudicing the result.

III. ELEMENTARY OPERATIONS AND CONSTRAINTS

3.1 PARALLEL AND SERIES REDUCTION; MINIMUM GRAPHS AND MINIMUM FORMS

3.1.1 Recognition of P/S Elements in T and associated N.C.'s:(a)

A. Write the t.s. form T of a graph G explicitly in terms of some two branch variables x and y,

$$T = T_{I} + xT_{II} + yT_{III} + xyT_{IV} \tag{3-1}$$

If $x /\!/ y$(b) in G, then $T_{I} \neq 0$, $T_{II} = T_{III}$ and $T_{IV} = 0$ (3-2)

If $x \not{S} y$(c) in G, then $T_{I} = 0$, $T_{II} = T_{III}$ and $T_{IV} \neq 0$. (3-3)

B. Associated N.C.: - Congruence with P./S. Manifestations — When a t.s.t. form is written after the manner of (3-1),

$$t = t_I + xt_{II} + yt_{III} + xyt_{IV} \tag{3-1a}$$

in order for $(t \rightarrow T)$(d):

if $t_{IV} = 0$, then it is necessary to have $t_I \neq 0$ and $t_{II} = t_{III}$, and in that case $x /\!/ y$ in the prospective graph of t;

if $t_I = 0$, then it is necessary to have $t_{IV} \neq 0$ and $t_{II} = t_{III}$, and in that case $x \varnothing y$ in the prospective graph of t.

C. Lemma to N.C. (3.1.1.B) — When a t.s.t. form is written after the manner of (3-1a), in order for $(t \rightarrow T)$, if $t_{II} = t_{III}$, then it is necessary to have either $t_I = 0$ or $t_{IV} = 0$, but not both.

3.1.2 P./S. Reduction and Minimum Resultants:

A. When a T or a t is found to contain $x /\!/ y$ (and/or $x \varnothing y$), by the criteria of 3.1.1, a parallel (and/or series) reduction of T or t can be effected by setting either $x = 0$ or $y = 0$ (and/or by differentiating it w.r. to x or y); e.g.,

$$t_{\substack{\text{parallel reduced} \\ \text{w.r. to y} \\ \text{(e)}}} \overset{(f)}{\triangleq} t_y = t\Big|_{y=0} = (t_I + xt_{II}) \text{ of (3-1a),}$$

$$t_{\substack{\text{series reduced} \\ \text{w.r. to y}}} \triangleq t_{y'} = \frac{\partial t}{\partial y} = (t_{III} + xt_{IV}) \text{ of (3-1a).}$$

B. A simple example will show that one stage of p./s. reduction of a graph may result in new sets of s./p. connected branches in the reduced graph (Figure III-1). An irreducible graph is obtained after the necessary successive p./s. reductions. Such an irreducible graph shall be referred to an a "minimum p./s. graph," or simply a "minimum graph."

One topological property of a minimum graph is that there are at least three branches incident on any one of its vertices.

A minimum T or t, symbolized by T_m or t_m, is characterized by ($T_{mI} \neq 0$ and $T_{mIV} \neq 0$) or ($t_{mI} \neq 0$ and $t_{mIV} \neq 0$) w.r. to any pair of its variables.

C. Associated N.C.: Inequality of t_{mII} and t_{mIII} — In order for ($t_m \to T_m$) it is necessary that $t_{mII} \neq t_{mIII}$ w.r. to any pair of x and y.

3.2 OPEN- AND SHORT-ELEMENT OPERATIONS; DISSECTION; REALIZABILITY OF D.P.'S

3.2.1 When a graph G is transformed by open- or short-element operation, its T is accordingly reduced to

$$T_{(\text{open-}x)} = T_x = T\Big|_{x=0} \tag{3-4}$$

or

$$T_{(\text{short-}x)} = T_{x'} = \frac{\partial T}{\partial x} . \tag{3-5}$$

Similarly, t_x $(=t\big|_{x=0})$ and $t_{x'}$ $(= \frac{\partial t}{\partial x})$ represent t with its x element opened and shorted, respectively, in its prospective graph.

3.2.2 Any t (or T) can always be partitioned into two sub-forms with respect to any arbitrary variable x in the following manner:

$$t = t_I + xt_{II} \, . \tag{3-6}$$

It is recognized that t_I and t_{II} are precisely t_x and t_x the open- and short-x reduced rorms of t respectively.

The process of decomposing a t into t_x and $t_{x'}$ for further separate treatment shall be called dissection of t with respect to x; t_x and $t_{x'}$ are the dissection products (d.p.'s).

3.2.3 Associated N.C.: -Universal Realizability of Dissection Products — In order for (t → T) it is necessary that its dissection products with respect to every one of its elements be realizable.

3.3 REVERSIBLE VS. IRREVERSIBLE OPERATIONS

It is to be noted that the parallel- and series-reductions are special cases of open- and short-element operations respectively. However, it is important to recognize that the former operations are homeomorphic(17) (thus topologically trivial) while the latter ones are not.

Any reduction operation on t, such as the parallel-series-reductions, is said to be a reversible operation in the sense that, between the original and the reduced forms, graphical realizability of the former implies that of the latter, and vice versa.

On the other hand, those reductions on t, such as the non-p./s. open- and short-element operations (where the realizability of the original form t implies that of the reduced form, but not vice versa), are said to be irreversibl

The origin of the difference between the above cases can be easily traced to the presence vs. the lack of certain relevant constraints in the t that is to be operated on.

3.4 COMPATIBILITY BETWEEN TWO GRAPHS, TWO t's AND THE TWO D.P.'S

Two labeled graphs containing the same branches and differing in their number of vertices by one are said to be compatible if, and only if, the lower rank graph is obtainable from the higher rank one by having a certain pair of its vertices to coalesce. This pair of vertices and the corresponding vertex in the lower rank graph are said to be the vertices of compatibility in the respective graphs. Figure III-2 gives an example in which graph (b) is redrawn isomorphically into (b') to bring its compatibility with graph (a) into better evidence.

It is to be noticed that labeling the branches in the graphs puts a severe requirement on the compatibility of these graphs. Thus, graphs (a) and (b) of the example are compatible only if their branches are appropriately labeled, or not labeled at all. (But un-labeled graphs are not of interest in our present study.)

Similarly, two T's (or t's) are said to be compatible if and only if their graphs (or their prospective graphs) are compatible.

One outstanding instance of compatibility between t's is given in the following: Associated N.C.: — Compatibility of Dissection Products — In order for (t → T) it is necessary that its dissection products w.r. to any one of its variables be compatible.

Compatibility between t.s. (or t.s.t.) forms is seen here defined in terms of their graphs. This indicates the fact that no simple algebraic criterion is readily available for its test, except in some elementary cases. As a matter of fact, to establish an indirect algebraic criterion for the compatibility between two realizable t.s. forms was the very center of interest in the original realizability study.

3.5 CUT-SETS AND RELATED TOPICS

3.5.1 Fundamental Cut-set $C(i/j)$ — A fundamental cut-set $C(i/j)$, associated with a tree $T^{(i)}$ and one of its branches $b^{(ij)}$, is obtainable from the t.s. form T by observing that every $x \varepsilon C(i/j)^{(g)}$ must form a tree with the 2-tree $T^{i\prime}_{j}$, in the complete tree set.

$[T^{i}_{j\prime} \triangleq T^{(i)}$ with $b^{(ij)}$ shorted $= \partial T^{(i)}/\partial b^{(ij)} = T^{(i)}/b^{(ij)}$

Therefore, set $C(i/j) = \{x \varepsilon\ [\partial T/\partial T^{i}_{j\prime}]\ \}^{(h)}$. (3-7)

3.5.2 Complete Set of Independent Cut-Sets $C^{(i)}$ — A collection of all fundamental cut-sets, $C(i/j)_{j=1,2,\ldots n}$, associated with $T^{(i)}$ constitutes a complete set of independent cut-sets pertaining to the tree $T^{(i)}$. It is designated by $C^{(i)}$. Similarly, a fundamental cut-set $C(i/j)$ and a complete set of independent cut-sets $C^{(i)}$ may be found for a t.s.t. form t.

3.5.3 Associated N.C.'s:

A. Vanishing of t upon Opening $\{x \varepsilon C(i/j)\}$ — In order for $(t \rightarrow T)$, it is necessary that any cut-set

C(i/j) of t, as determined after Article 3.5.1, must satisfy

$$t\Big|_{\{x\}=0} = 0, \ \{x \varepsilon C(i/j)\} \qquad (3\text{-}8)$$

B. Inclusiveness of All Variables in $C^{(i)}$ — In order for (t → T), it is necessary that the set of elements in $C^{(i)}$ be inclusive of all variables of t.

3.5.4 Adjoint and Disjoint Cut-sets; Pencil-sets: — To save space, the definitions of these sets shall be relegated to the graphical clarification in Figure III-3, although the term pencil-set is in common use. The following pertinent remarks are made here:

(a) With respect to any conceived pair of vertices (V_x, V_y), an independent set of cut-sets $C^{(i)}$ pertaining to any tree $t^{(i)}$ must contain k disjoint cut-sets, where k is the "path length" between V_x and V_y in $t^{(i)}$, and is $\geqq 1$.

(b) Any set $C^{(i)}$ must contain m pencil-sets, where m is the number of tips in the tree $t^{(i)}$, and $m \geqq 2$.

3.6 CUT-REMAINDER; ITS INVARIANCE OVER {xεC(i/j)}

3.6.1 Graphically, the cut-remainder of a graph G consists of two separate subgraphs resulting from a cutting operation on G. In order to make its t.s. form representation possible, one may consider the "re-attached cut-remainder graph." Thus, the cut-remainder form, $\hat{t}(i/j)$, pertaining to a cut C(i/j) of a given t may be obtained by first shorting out the branch $b^{(ij)}$ and then opening the rest of $\{x \varepsilon C(i/j)\}$, i.e.,

$$\hat{t}(i/j) = \frac{\partial t}{\partial b^{(ij)}}\Bigg|_{\{x\}=0} \triangleq t_{(ij)'}\Bigg|_{\{x\}=0}$$

$$\{x \varepsilon C(i/j)\} \ . \qquad (3\text{-}9)$$

3.6.2 Associated N.C.: Invariance of Cut-remainder w.r. to $\{x\varepsilon C(i/j)\}$ — In order for $(t \rightarrow T)$, it is necessary that any cut-remainder of t must be invariant w.r. to all elements of the corresponding cut-set. Thus, considering (3-8) and (3-9), one has for every $C(i/j)$,

$$t = [0] + (\Sigma x)^{\cdot}\hat{t}(i/j) + \begin{bmatrix}\text{Terms containing more}\\ \text{than one factors from}\end{bmatrix}\{x\} ,$$

$$\{x\varepsilon C(i/j)\} \ . \qquad (3\text{-}10)$$

3.7 LOOP-SETS AND RELATED TOPICS**

3.7.1 Fundamental Loop Set $L(i/k)$ — A fundamental loop-set $L(i/k)$, associated with a tree $T^{(i)}$ and one of its cords $b^{(ik)}$, is obtainable from the t.s. form T by opening all cords of $T^{(i)}$ except $b^{(ik)}$. The resultant sub-form $T]L(i/k)]$ will be

$$T[L(i/k)] = \left[a_1 a_2 \ldots . a_{(n-\lambda+)} l_1 l_2 l_3 \ldots . l_\lambda \left(\frac{1}{l_1} + \frac{1}{l_2} + \frac{1}{l_3} + \ldots\ldots + \frac{1}{l_\lambda}\right)\right]^{i}_{k} \qquad (3\text{-}11)$$

where l's are the loop elements, in particular, l_1 to $l_{(\lambda-1)}$ belong to $t^{(i)}$ and l_λ is the cord $b^{(ik)}$, a's are the appended i-th tree elements, λ is the size of the loop and n, the order of T. The set $L(i/k)$, then, contains all the elements in the cyclic part of $T[L(i/k)]$, denoted as

$$\text{Set } L(i/k) - \{x\varepsilon T_{\circlearrowright}[L(i/k)]\} \ . \qquad (3\text{-}12)$$

3.7.2 Associated N.C.: Minimum Allowable Size of Loop from A t_m — In order for ($t_m \rightarrow T_m$), it is necessary that $\lambda \geqq 3$.

3.7.3 Complete Set of Independent Loop-sets $L^{(i)}$ — A collection of all fundamental loop-sets, $L(i/k)_{k=1,2....(B-n)}$, associated with $T^{(i)}$ constitutes a complete set of independent loop-sets $L^{(i)}$.

3.7.4 λ-tree-sum Form from Shrinking A Loop; Its Invariance over $\{x \varepsilon L(i/k)\}$ —

A. The λ-tree-sum form, $t(\lambda|i/k)$, from shrinking a loop $L(i/k)$ of size λ, is obtained from t by shorting out any (λ-1) of the λ elements of the pertinent loop. It is readily expressible in terms of one particular set of the loop elements, namely,

$$t(\lambda|i/k) = \partial^{(\lambda-1)} t / \partial l_1 \partial l_2 \ldots . \partial l_{(\lambda-1)} . \qquad (3\text{-}13)$$

B. Association N.C.: Invariance of $t(\lambda|i/k)$ over $\{x \varepsilon L(i/k)\}$ — In order for ($t \rightarrow T$), it is necessary that the λ-tree-sum forms resulted from shrinking of the loop $L(i/k)$ must be invariant w.r. to any one loop element. Thus, considering (3-11) and (3-13), one has for every $L(i/k)$,

$$t = [0] + \left[l_1 l_2 \ldots . l_\lambda \left(\frac{1}{l_1} + \frac{1}{l_2} + \ldots\ldots + \frac{1}{l_\lambda} \right) \right]^i_k .$$

$$t(\lambda|i/k) + \begin{bmatrix} \text{Terms containing less than} \\ (\lambda\text{-1}) \text{ factors from } \{x\} \end{bmatrix} ,$$

$$x \varepsilon L(i/k). \qquad (3\text{-}14)$$

IV. HIGHER ORDER OPERATIONS AND CONSTRAINTS

4.1 DEGENERATE GRAPHS AND FORMS: FACTORIZATION OF A DEGENERATE T.S.T. FORM

4.1.1 Definition and Related Topics:

A. A graph in which two or more inseparable sub-graphs are chain- or tree-connected is a degenerate graph; it is not uniquely definable, but is one of a set of the 1-isomorphic graphs.(18) The t.s. form of such a degenerate graph is known to be the product of the t.s. forms of its constituent sub-graphs.

$$T = {}^{(1)}T \cdot {}^{(2)}T \cdot {}^{(3)}T \cdots \qquad (4\text{-}1)$$

B. A Necessary and Sufficient Condition — A degenerate t,

$$t = {}^{(1)}t \cdot {}^{(2)}t \cdot {}^{(3)}t \qquad (4\text{-}1a)$$

is realizable as the T of any one of a set of 1-isomorphic graphs if, and only if, each one of its factors is individually realizable.

C. A degenerate graph, or t, may come into a problem as a given entity, or it may be produced through an open-element, a short-element, or a cut operation (Figure IV-1(a)(b)(c)).

4.1.2 Factorization of A T.S.T. Form — Factorization is a very desirable algebraic decomposition of a t.s.t. form whenever possible, as it will transform a high order problem into two or more much lower order problems. It is also an important auxiliary operation and knowledge to certain other operations (Articles 4.2.1 and 4.5.1).

Only the factorization of a minimum t is of interest.

Thus, a non-trivial case would involve a t of at least 6th order in no less than 12 variables.

A. Principle for Determining Factorizability —

Graphical observations (Figure IV-2) — Any fundamental cut-set of a degenerate graph consists of elements from one and only one of its constituent subgraphs. Hence, a complete set of independent cut-sets $C^{(i)}$ of such a graph is decomposable into as many disjoint sub-sets as there are constituent subgraphs.

Algebraic observation — Because of the homogeneity and multi-linearity in its variables, any factorizable t must have its factors homogeneous in, and mutually exclusive of, their respective groups of variables. This implies that each term in such a t must contain a fixed number of variables from each factor, which in turn implies that the cut operation $\partial t/\partial[t^{(i)}/b^{(ij)}]$ can only yield elements from the factor to which $b^{(ij)}$ belongs. Furthermore, as any $C^{(i)}$ is inclusive of all variables of t, the entire set of variables must be exposed in separable groups through $C^{(i)}$ if t is factorizable. Therefore, one has:

B. The Procedure for Determining the Factorizability of t^{**} —

(i) Find the set of cut-sets $C^{(i)}$ associated with any arbitrary tree $t^{(i)}$

(ii) Determine the separability among the cut-sets

$C(i/j)$ of $C^{(i)}$ by writing down the cut-sets vs. branch-element incident matrix $|C^{(i)}|$ in an obvious manner.

The result is a cut-set vs. branch-element incident matrix with its separability, if any, placed in evidence.

Each disjoint group in $|C^{(i)}|$ ought to yield a factor in t. The actual factorization of t is but a simple matter of sorting, of which the success is guaranteed by the following N.C. to the extent that t is possibly realizable.

C. Associated N.C.: Factorizability consistancy condition — In order for $(t \rightarrow T)$, it is necessary that t be actually factorizable into as many factors as may be revealed by the disjoint groupings in $|C^{(i)}|$.

4.2 FRAGMENTS IN A CUT-REMAINDER; TRIMMED GRAPHS AND FORMS; $C(i/j) - F(i/j)$ INCIDENCE RELATIONS

4.2.1 Fragments in A Cut-remainder — A cut-remainder form $\hat{t}(i/j)$ is very commonly factorizable, i.e.,

$$\hat{t}(i/j) = \Big[{}^{(1)}f \cdot {}^{(2)}f \cdot {}^{(3)}f \ldots\ldots \Big]_j^i \quad . \qquad (4\text{-}2)$$

These factors, and their corresponding prospective graphs, shall be referred to as the fragments from the $C(i/j)$ cut (Figure IV-1(c)). The fragment-set is denoted by $f(i/j)$, with its individual members denoted by ${}^{(k)}f(i/j)$.

4.2.2 Trimmed Graphs and Forms —

A. A trimmed graph $\dot{G}(i/j)$ is one resulting from opening all elements in G except those of $t^{(i)}$ and $C(i/j)$ (Figure IV-3(a)). Similarly, a trimmed t.s.t. form $\dot{t}(i/j)$ is obtained as

$$\dot{t}(i/j) \triangleq t \left| \begin{array}{l} \{x\} \neq 0,\ x \varepsilon C(i/j) \\ \{y\} \neq 0,\ y \varepsilon t^{(i)} \\ \text{All other elements} = 0. \end{array} \right. \tag{4-3}$$

B. A doubly trimmed graph $\ddot{G}(i/j)$ is one obtained from $\dot{G}(i/j)$ by shorting out all of its free branches and by reducing the series-connected $t^{(i)}$-elements within each $\dot{F}(i/j)$ (Figure IV-3(b)).

In the extreme case, an entire fragment may thus be trimmed to a single element.

It is to be noted that the doubly trimmed graph retains the entire incidence relationship between the members of the fragment-set and those of the cut-set.

4.2.3 Simple $C(i/j)$ - $F(i/j)$ Incidence — A complete incidence relationship between the cut-set elements and the fragments may not be simply explored due to complications raised by multiple incidence. On the other hand, knowledge of simple incidence of cut-set elements on the fragments can be easily obtained, and it offers a specific use as will be seen in the next section.

By simple incidence here it is meant that single cut-set elements are incident on individual exclusive vertices of a certain fragment. This condition is tested out when that entire fragment is open circuited. For then all these simple incident cut-set elements will become degenerate branches in the testing subgraph (Figure IV-4(a)) and are exhibited as free factors in the testing sub-form of t, or that of $\ddot{t}(i/j)$.

It is to be noted that, under certain conditions, some of the free factor elements thus exposed may actually be incident on some neighboring fragments to the one under consideration, as shown shaded in Figure IV-4(b). However, such an infringing element is easily identified when the shaded fragment has its turn to be similarly explored.

4.3 TURNABLE SUBGRAPH AND ITS CONDENSATION; ESSENTIAL-MINIMUM GRAPHS AND FORMS; CONDENSATION OF A T.S.T. FORM

4.3.1 Definition and Related Topics:

A. A subgraph that is connected to the rest of the graph at two and only two meshing vertices is called a turnable subgraph; it can be "turned around,"(18) so to speak, across these meshing vertices, since its orientation w.r. to the main graph is of no significance in the present study.

Turnable subgraphs may be a given entity or it may be produced as a result of open- and short-element operations on a graph (Figures IV-5(a) and (b)).

Without losing generality, let the case of a graph consisting of two turnable subgraphs $\bar{G}$ and $\bar{\bar{G}}$, meshed at vertices V_A and V_B, be considered (Figures IV-5(a) and (b)). It is well known(1) that the t.s. form of the total graph G may be expressed as:

$$T = \bar{T} \cdot \bar{\bar{T}}' + \bar{T}' \cdot \bar{\bar{T}} \qquad (4\text{-}4)$$

where $\bar{T}$ and $\bar{\bar{T}}$ are the t.s. forms of $\bar{G}$ and $\bar{\bar{G}}$, $\bar{T}'$ and $\bar{\bar{T}}'$ are the 2-t.s. forms of $\bar{G}$ and $\bar{\bar{G}}$ w.r. to (V_A, V_B) and $\bar{T}$, $\bar{T}'$, $\bar{\bar{T}}$, and $\bar{\bar{T}}'$ are respectively of orders $\bar{n}$, $(\bar{n}-1)$, $\bar{\bar{n}}$, and $(\bar{\bar{n}}-1)$.

B. Any turnable subgraph can be condensed into a single branch without altering the topological properties of the companion subgraph. Thus, the Graph G is said to be condensible into $\tilde{G}$ and $\tilde{\tilde{G}}$ as shown in Figure IV-6. The corresponding t.s. forms become:

$$\tilde{T} = \bar{T} + \bar{\bar{b}}_o \bar{T}'$$

$$\text{and} \quad \tilde{\tilde{T}} = \bar{\bar{T}} + \bar{b}_o \bar{\bar{T}}' \tag{4-5}$$

where $\bar{\bar{b}}_o$ and $\bar{b}_o$ are single branch substitutes of $\bar{\bar{G}}$ and $\bar{G}$ respectively.

C. A minimum graph that is non-condensible shall be called an essential-minimum graph. Similarly, a non-condensible minimum t.s., or t.s.t., form shall be called an essential-minimum form, represented by T_M, or t_M.

4.3.2 Condensation of T.S.T. Forms and Related Topics:

A. A condensible t.s.t. form is defined purely on the basis of its algebraic formal structure with no regard to the realizability of its factored sub-forms and to their compatibility. That is, t is said to be condensible if and only if

$$t = \bar{t}\ \bar{\bar{t}}' + \bar{t}'\ \bar{\bar{t}} \tag{4-4a}$$

where $\bar{t}$, $\bar{t}'$, $\bar{\bar{t}}$, and $\bar{\bar{t}}'$ are respectively of orders $\bar{n}$, $(\bar{n}-1)$, $\bar{\bar{n}}$, and $(\bar{\bar{n}}-1)$ in the sets of variables $\{\bar{x}\}$, $\{\bar{x}'\}$, $\{\bar{\bar{x}}\}$, and $\{\bar{\bar{x}}'\}$,

in which $\{\bar{x}\} = \{\bar{x}'\}$ and $\{\bar{\bar{x}}\} = \{\bar{\bar{x}}'\}$

and $\{\bar{x}\} \cap \{\bar{\bar{x}}\} = 0$.

B. "Termwise manifestation of condensibility" of t refers to the occurrence when, by some yet to be explained procedures, the two sub-sets of elements $\{\overline{x}\}$ and $\{\overline{\overline{x}}\}$ can be identified and all terms in t are found to be of adequate orders in $\{\overline{x}\}$ and $\{\overline{\overline{x}}\}$ and so that the two proper partial sums can be exhibited. It can be proved that this manifestation signifies the condition that there are two and only two vertices in the prospective graph of t on which some elements from both $\{\overline{x}\}$ and $\{\overline{\overline{x}}\}$ have common incidence. As this topological argument establishes a foregone conclusion for the existence of a turnable subgraph for such a t, one has:

C. An Associated N.C.: Consistancy of t with its Term-wise Manifestation of Condensibility — Whenever all terms in a given t are found to manifest condensibility for t, then in order for $(t \rightarrow T)$, it is necessary that the pertinent partial sums be properly factorizable so as to be consistant with the manifestation.

D. To effect condensation of t from (4-4a), one merely writes the doublets:

$$\left.\begin{array}{ll} & \tilde{t} \overset{\Delta}{=} \begin{pmatrix} \text{t with } \overline{\overline{t}} \text{ condensed} \\ \text{into a single branch } \overline{\overline{b}}_o \end{pmatrix} = \overline{t} + \overline{\overline{b}}_o \overline{t}' \\ \text{and} & \tilde{\tilde{t}} \overset{\Delta}{=} \begin{pmatrix} \text{t with } \overline{t} \text{ condensed} \\ \text{into a single branch } \overline{b}_o \end{pmatrix} = \overline{\overline{t}} + \overline{b}_o \overline{\overline{t}}' \end{array}\right\} \quad (4\text{-}5a)$$

Like factorization, condensation transforms a high order problem to problems of much lower orders. It is also a necessary reduction procedure in dealing with certain problems.

E. Associated Necessary and Sufficient Condition — If t is condensible, then in order for $(t \to T)$, it is necessary and sufficient that $\tilde{t}$ and $\tilde{\tilde{t}}$, the condensed doublets, be realizable individually.

4.4 EXAMINATION OF A T.S.T. FORM FOR CONDENSIBILITY

Only the examination of a minimum t_m (or T_m) is of practical interest. The smallest condensible minimum graph is shown in Figure IV-7. Hence any minimum t_m (or T_m) less than 5th order or having fewer than ten variables cannot be condensible.

4.4.1 Principles for Uncovering A Turnable Sub-set $\{\overline{x}\}$:

A. Direct Examination — To save space, only the following remarks on this method will be made:

(1) What is recommended is a part of a complete examination scheme. Because of its straightforwardness and simplicity, it is offered for optional use, particularly under certain circumstances as will be commented on shortly.

(2) The structure of a graph (or a t.s.t. form t) is examined based on knowledges revealed through a fundamental cut $C(i/j)$. The situation of a $\overline{G}$ w.r. to the cut is the object of observation. The four possible cases are shown in Figure IV-8.

(3) The use of doubly trimmed forms $\ddot{t}(i/j)$ greatly facilitates the algebraic examination of t.

(4) Referring to Figure IV-8, $\overline{G}$ in case (c) is most readily revealed by the occurrence of a trivial

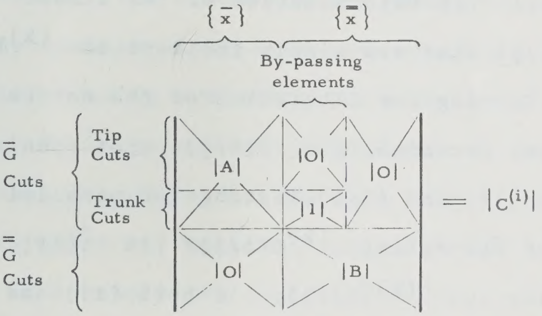
fragment in $\ddot{t}(i/j)$. $\overline{G}$ in case (a) is also easily revealed by a test designed around the knowledge of simple $C(i/j)$ - $F(i/j)$ incidence relations.

(5) Case(d) of Figure IV-8 is not amenable to treatment for the purpose of tracing $\overline{G}$. But, $\overline{G}$ can not remain in obscurity under the cover of this case over the complete set of cuts $C^{(i)}$.

(6) A procedure has been developed for case(b); but it is not recommended because of its elaborateness. As the occurrence of cases (a) and (c) cannot be guaranteed to show up in the set of cuts $C^{(i)}$, this method is left incomplete by avoiding treatment of case (b).

B. Indirect Examination**

Let the contents of a complete set of cut-sets $C^{(i)}$ be examined. Without losing generality, a typical case is shown, in Figure IV-9, of a graph containing a turnable subgraph $\overline{G}$. An arbitrarily chosen tree $T^{(i)}$ is shown in heavy lines. It is seen that, if not for the presence of the dashed elements of $\overline{\overline{G}}$, one would have a degenerate graph consisting of at least two hinged subgraphs of which $\overline{G}$ (in its entirety) is one. The set of cuts $C^{(i)*}$ from such a depleted graph would be separable in a cut vs. branch matrix. Now, when the dashed elements are restored, the $\{\overline{\overline{x}}\}$ part of the above depleted matrix is augmented by inclusion of these elements; at the same time, the separability of the depleted matrix is spoiled. However, one may observe that, wherever the $\overline{\overline{G}}$ elements are coupled with $\overline{G}$ elements in the said matrix through one or more $\overline{G}$ cuts it is always the entire sub-set of the dashed elements of $\overline{\overline{G}}$ that provides the coupling. This is strictly due to the two-vertex connectedness of $\overline{G}$ with $\overline{\overline{G}}$ in the total graph G — Any $\overline{\overline{G}}$ element(s) that gets in the way of any trunk cut, so to speak, of $\overline{G}$ must be the one(s) that completely by-passes $\overline{G}$ in G, and therefore will also get in the way of all trunk cuts of $\overline{G}$. The pertinent matrix structure thus appears as:


$\{\overline{x}\}$
$\{\overline{\overline{x}}\}$
By-passing elements
$\overline{G}$ Cuts
Tip Cuts
Trunk Cuts
$\overline{\overline{G}}$ Cuts
|A|
|O|
|O|
|1|
|O|
|B|
= $|C^{(i)}|$

(4-6)

where the sub-matrices |A| and |B| are ordinary, |O|'s are empty and |1| is an entire matrix in which all entries are 1's.

It is the author's experience that the above pattern of the sub-matrices, if it should exist at all, can be placed into evidence with little difficulty. In most non-trivial cases, the relatively large counts of 1's entries in the trunk-cut rows of $\overline{G}$ and in the by-passing element columns of $\overline{\overline{G}}$ help to identify these rows and columns, and considerably facilitate the rearrangement of $|C^{(i)}|$ into the desired pattern. Thus, one has:

4.4.2 Procedures for uncovering $\{\overline{x}\}$ from a T.S.T. Form t — The cut-set methods:

A. The Incomplete Direct Procedure — In the interest of brevity, the detailed steps of this procedure will be left to the reader. By referring to the remarks (3), (4), and (5) of Article 4.4.1(A), one should find no difficulty in establishing these steps except, perhaps, those for the case (a) of Figure IV-8. They involve: (i) Determination of the subset $C(i/j)^*$ of $C(i/j)$ that are simply incident on ${}^{(k)}f(i/j)$; (ii) Testing for liberation of the entire ${}^{(k)}\ddot{f}(i/j)$ (as free factors) from $\ddot{t}(i/j)$ upon opening ${}^{(k)}C(i/j)^*$, and (iii) Testing for parallel connectedness of the entire ${}^{(k)}C(i/j)^*$ in $\ddot{t}(i/j)$ upon shorting out ${}^{(k)}\ddot{f}(i/j)$. If both (ii) and (iii) tests are successful, then case (a) of Figure IV-8 exists and

$$\{\overline{x}\} = \{{}^{(k)}C(i/j)^*\} \bigcup \{{}^{(k)}f(i/j\} \ . \qquad (4\text{-}7)$$

B. The Indirect and Conclusive Procedure —

1. Take an arbitrary tree $t^{(i)}$ from t.
2. Obtain the complete set of cut-sets $C^{(i)}$.
3. Write down the $|C^{(i)}|$ matrix.
4. Examine $|C^{(i)}|$ for possible rearrangement of its rows and columns so as to exhibit the pattern of (4-6

If 4. is successful, $\{\bar{x}\}$ is identified.

If 4. is not successful, t is non-condensible.

C. Comment on the Procedures —

Procedure B is most likely simpler to apply than procedure A. Its only drawback is the seeming presence of an uncertainty about the examiner's capability of securing the desired matrix pattern when no success is reported. This is of theoretical concern, especially when a multiplicity of turnable subgraphs or compound turnable subgraphs exist. Nevertheless, from a practical point of view, such uncertainty is rather remote even in the solution of problems of considerable sizes.

For all the elaborateness needed for achieving its already limited objectives, procedure A possesses the benefit of being definitive in every step of its operations. It is more readily adaptable to machine execution than procedure B. For this reason, it is included in this paper for optional use, although it cannot serve as a conclusive negative test.

In any event, positive findings of $\{\bar{x}\}$ through the use of either procedure can be proved to be a definite consequence of its being a turnable sub-set in the prospective graph of t, hence, it requires:

D. Associated N.C.: - Consistancy of t with the Findings of $\{\bar{x}\}$ — When a turnable sub-set $\{\bar{x}\}$ is yielded by the cut-set methods of Art. 4.4.2 for a given t, then in order for $(t \rightarrow T)$, it is necessary that t be actually condensible.

4.5 PENCIL-SETS IN AN ESSENTIAL MINIMUM GRAPH; A PRACTICAL REALIZATION EXAMINATION OF t_M THROUGH TRIAL**

4.5.1 An essential minimum graph G_M has the unique property that the cut-remainder of any one of its entire set of pencil-sets must be a non-degenerate graph. This can be seen plainly from Figure IV-10, in which the illustrated graph could not have been an essential-minimum one if any one of its pencil-cut-remainders is degenerate.

This unique property of a G_M, or t_M, enables one to identify its pencil-set(s) by the rule:

> A cut-set of an essential minimum G_M or t_M is a pencil-set when and only when its cut-remainder is non-degenerate.

4.5.2 With the above simple rule for identifying pencils among cut-sets of an essential-minimum t.s.t. form t_M, and also with the knowledge that any complete set of independent cut-sets must contain at least two pencil sets, one may conceive of the practicality of uncovering all the pencils of a t_M by a number of trials, which may be conducted in a certain orderly manner to effect a fast result.

In fact, in a practical problem, such a trial approach may not prove to be unadvisable at all in arriving at a quick answer to the realizability of t_M. Hence, having uncovered all pencils of t_M, consistancy among them is first established by the construction of the graph G_M, from which a T_M can be written out and checked against t_M term by term.

This rather efficient approach to the solution of realizability problems, in which specific unknowns are given, unfortunately has not been made serviceable in the theoretical analysis of an arbitrary t.s.t. form. Through the use of pencil sets in a t_M, there ought to be some hope for the establishment of a truly powerful high order constraint in a realizable T_M. Perhaps, this will make for an interesting quest.

V. MISCELLANEOUS CONSTRAINTS

The following constraints are very convenient means for expeditious disposal of un-resolved t.s.t. forms. Because of lack of space, only their titles are given here to indicate the nature of the constraints. An interested reader may refer to the original dissertation.

5.1 NUMBER OF SERIES-PAIRS PRODUCED UPON OPENING A SINGLE ELEMENT FROM A MINIMUM T.S.T. FORM, N^{g}

5.2 ADMISSIBLE SET SIZE, B(n), FOR AN ^{n}TH ORDER t_m

5.3 ADMISSIBLE SIZES, ν (n,B), FOR A GIVEN $t_m(n,B)$

5.4 CONDENSIBILITY OF $t_m(n,B)$ OF CERTAIN SIZES

5.5 REALIZABLE NON-PARALLEL 3-RD ORDER T.S.T. FORMS

VI. CONCLUSION

A large collection of algebraic tools is made available for the treatment of the proposed realization problem. The many structural constraints in t.s. forms thus far exposed are hoped to be of enough interest to elicit further quest for more crucial conditions. It is also hoped that pursuit of knowledge in this area may yield some useful analytical insight apparently desired in expediting a solution for the n-port network synthesis problem as voiced by several proponents in this field.(19)

VII. ACKNOWLEDGMENT

The author wishes to use this opportunity to gratefully thank Prof. R. M. Foster for his inspiring guidance and encouragement throughout the thesis work, from part of which this material is taken.

Abbreviations and Symbols:

(a)	N.C. stands for	"Necessary Condition"
(b)	// stands for	"is parallel connected with"
(c)	$\not{S}$ stands for	"is series connected with"
(d)	$(t \to T)$ stands for	"t to be realizable as a T"
(e)	"w.r. to"	"with respect to"
(f)	$\underline{\triangle}$ signifies	"equal by definition"
(g)	$x \varepsilon X$ signifies	"x that is an element of X"
(h)	$\{x \varepsilon X\}$ signifies	"entire set of x that are elements of X"

BIBLIOGRAPHY

1. W. S. Percival, "The Solution of Passive Electrical Networks by Means of Mathematical Trees" Proc. IEEE, Vol. 100, Pt. III (May 1953), pp. 143-150.

2. Y. H. Ku, "Resume of Maxwell's and Kirchhoff's Rules for Network Analysis" J. Franklin Inst., Vol. 253, No. 3 (March 1952), pp. 211-224.

3. L. Weinberg, "Kirchhoff's Third and Fourth Law's," IRE Transaction on Circuit Theory, Vol. CT-5, No. 1 (March 1958), pp. 8-30.

4. H. M. Trent, "A Note on the Enumeration and Listing of All Possible Trees in A Connected Linear Graph," Proc. National Academic Science, Vol. 40 (October 1954), pp. 1004-1007.

5. N. Nakagawa, "On Evaluation of the Graph Trees and the Driving Point Admittance," IRE Transactions on Circuit Theory, Vol. CT-5, No. 2 (June 1958), pp. 122-127.

6. S. C. Lee, "On Topological Formulae," Proc. of First Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois (1963), pp. 427-455.

7. S. D. Bedrosian, "Evaluation of Network Determinants via the Key Subgraphs," Proc. of First Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois (1963), pp. 589-602.

8. S. L. Hakimi and D. G. Green, "Generation and Realization of Trees and k-trees," IEEE Transactions on Circuit Theory, Vol. CT-11, No. 2 (June 1964), pp. 247-255.

9. R. M. Foster, "Topologic and Algebraic Considerations in Network Synthesis," Proc. of the Symposium on Modern Network Synthesis, Vol. I (1952), pp. 8-18.

10. S. Okada, "Topology Applied to Switching Circuits," Proc. Symposium on Information Networks, Polytechnic Inst. of Brooklyn (April 1954), pp. 267-290.

11. S. Seshu, "Topological Considerations in the Design of Driving Point Functions," IRE Transaction on Circuit Theory, Vol. CT-2, No. 4 (December 1955), pp. 356-367.

12. R. Gould, "Graphs and Vector Spaces," J. Math & Physics, Vol. 37 (1958), pp. 193-214.

13. E. A. Guillemin, "How to Grow Your Own Trees From a Given Cut-set or Tie-set Matrices," IRE Transaction on Circuit Theory, Vol. CT-6, Special Supplement (May 1959), pp. 110-126.

14. L. Lofgren, "Irredundant and Redundant Boolean Branch-Networks," IRE Transaction on Circuit Theory, Vol. CT-6, Special Supplement (May 1959), pp. 158-175.

15. O. Wing and W. H. Kim, "The Path Matrix and Its Realizability," IRE Transactions on Circuit Theory, Vol. CT-6, No. 3 (September 1959), pp. 267-272.

16. W. Mayeda, "Necessary and Sufficient Conditions for Realizability of Cut-set Matrices," IRE Trans. on Circuit Theory, Vol. CT-7, No. 1 (March 1960), pp. 79-81

17. H. Whitney, "A Set of Topological Invariants for Graphs," Amer. J. of Math, Vol. 55 (1933), pp. 231-235.

18. H. Whitney, "On the Classification of Graphs," Amer. J. of Math, Vol. 55 (1933), pp. 236-244.

19. "The Realization of n-port Networks without Transformers--A Panel Discussion," IRE Transactions on Circuit Theory, Vol. CT-9, No. 3 (September 1962), pp. 202-214.

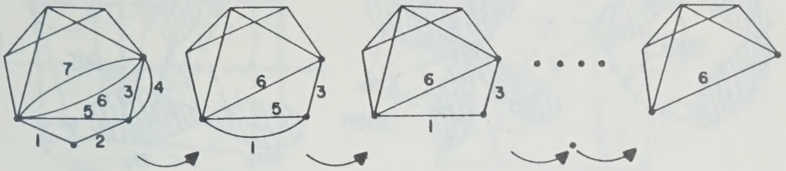
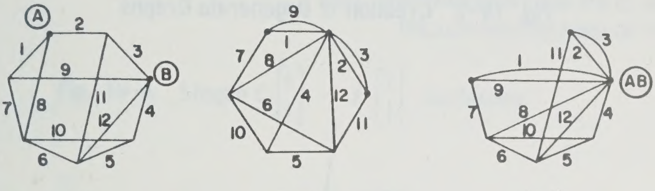

7
5
6
3
4
1
2
6
5
3
1
6
3
1
6

Fig. III-1. Production of a Min. Graph following Successive P./S. Reduction


A
2
1
3
9
B
7
8
11
12
4
10
6
5
9
7
8
1
3
2
6
4
12
11
10
5
3
11
2
1
9
AB
7
8
10
12
4
6
5

(a) (b) (b')

Fig. III-2. Example of Compatible Graphs

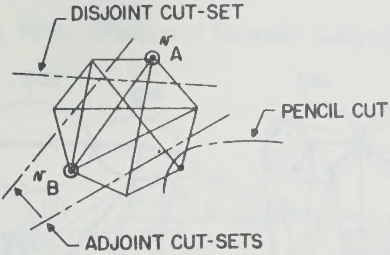

DISJOINT CUT-SET
A
PENCIL CUT
B
ADJOINT CUT-SETS

Fig. III-3. Disjoint and Adjoint Cuts; Pencil Cut

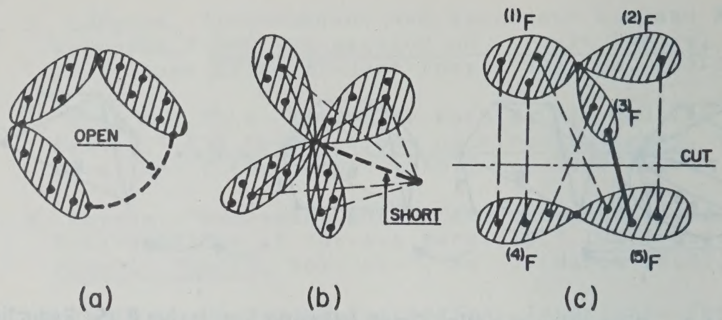

$^{(1)}F$
$^{(2)}F$
$^{(3)}F$
OPEN
SHORT
CUT
$^{(4)}F$
$^{(5)}F$
(a)
(b)
(c)

Fig. IV-1. Creation of Degenerate Graphs

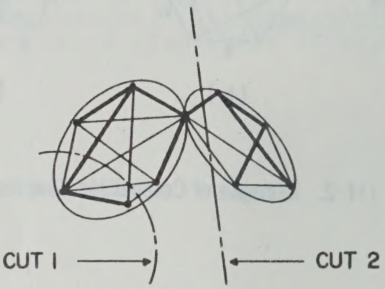


Fig. IV-2. Separability of Cut-sets in a Degenerate Graph

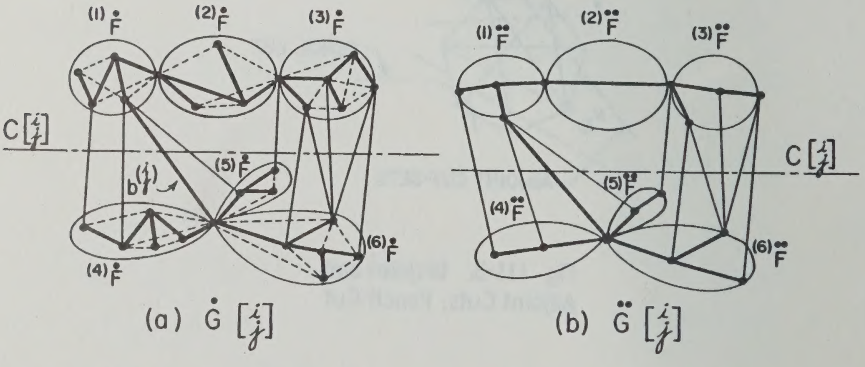
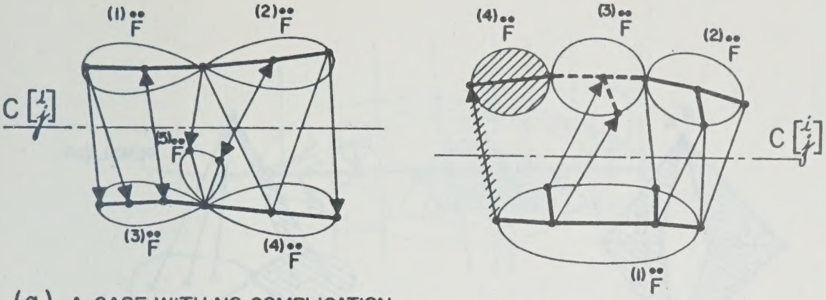

$^{(1)}\dot{F}$
$^{(2)}\dot{F}$
$^{(3)}\dot{F}$
$C\left[\frac{i}{j}\right]$
$b^{(\frac{i}{j})}$
$^{(5)}\dot{F}$
$^{(4)}\dot{F}$
$^{(6)}\dot{F}$
(a) $\dot{G}\left[\frac{i}{j}\right]$
$^{(1)}\ddot{F}$
$^{(2)}\ddot{F}$
$^{(3)}\ddot{F}$
$C\left[\frac{i}{j}\right]$
$^{(5)}\ddot{F}$
$^{(4)}\ddot{F}$
$^{(6)}\ddot{F}$
(b) $\ddot{G}\left[\frac{i}{j}\right]$

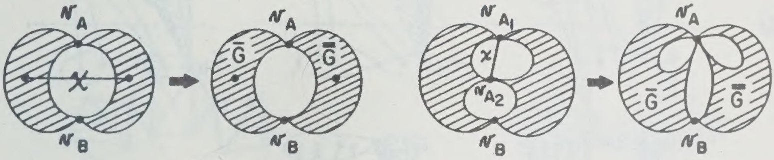
Fig. IV-3. (a) Trimmed Graph; (b) Doubly Trimmed Graph


(1) F
(2) F
C [i/j]
(5) F
(3) F
(4) F
(4) F
(3) F
(2) F
C [i/j]
(1) F

(a) A CASE WITH NO COMPLICATION.

(b) INFRINGEMENT AND ITS CLARIFICATION FOLLOWING OPENING OF $^{(3)}\ddot{F}$.

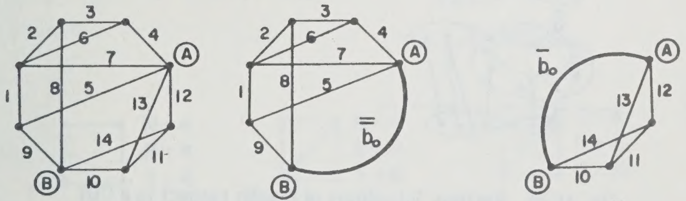
Fig. IV-4. Simple $C\begin{bmatrix} i \\ j \end{bmatrix}$ - $F\begin{bmatrix} i \\ j \end{bmatrix}$ Incidence


N A
x
N B
N A
G
G
N B
N A1
x
N A2
N B
N A
G
G
N B

(a) BY OPENING x

(b) BY SHORTING x

Fig. IV-5. Creation of Turnable Subgraphs


3
2
6
4
A
7
1
8
5
13
12
14
9
11
B
10
3
2
6
4
A
7
1
8
5
9
b₀
B
b₀
A
13
12
14
11
B
10

(a) G

(b) $\widetilde{G}$

(c) $\overset{\approx}{G}$

Fig. IV-6. Transformation of G into $\widetilde{G}$ and $\overset{\approx}{G}$ by Condensation

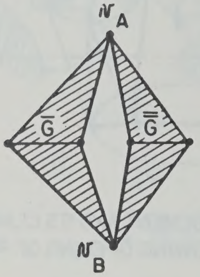

$\mathcal{N}_A$
$\overline{G}$
$\overline{\overline{G}}$
$\mathcal{N}_B$

Fig. IV-7. Smallest Condensible Minimum Graph

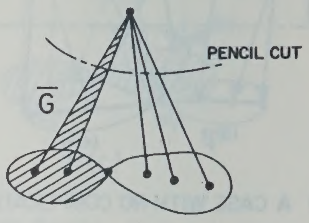

PENCIL CUT
$\overline{G}$

Fig. IV-10. A Pencil Cut with Degenerate Cut Remainder

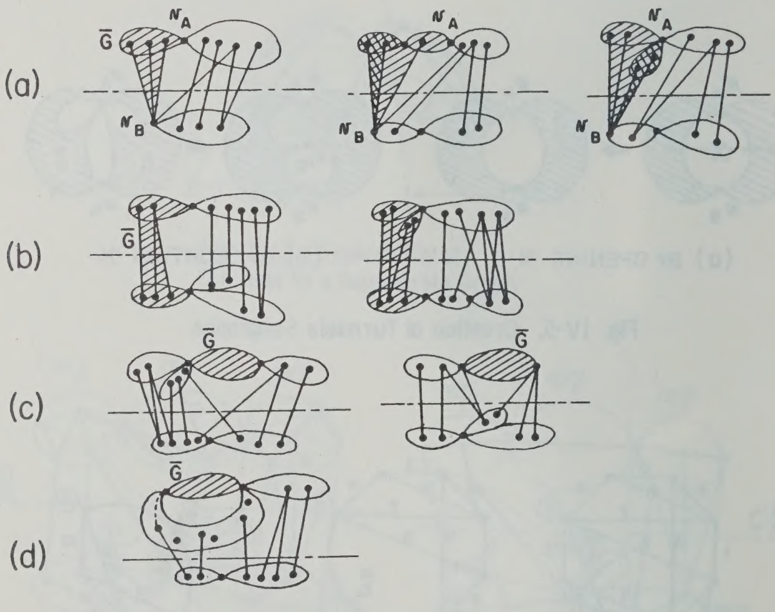

(a)
(b)
(c)
(d)
$\mathcal{N}_A$
$\mathcal{N}_B$
$\overline{G}$

Fig. IV-8. Various Situations of $\overline{G}$ with respect to a Cut

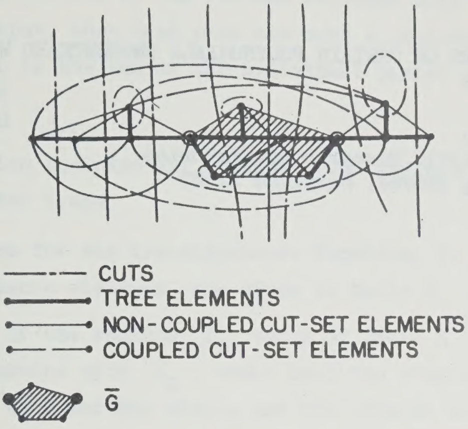

CUTS
TREE ELEMENTS
NON-COUPLED CUT-SET ELEMENTS
COUPLED CUT-SET ELEMENTS
$\bar{G}$

Fig. IV-9. Examination of a Complete Set of Cuts

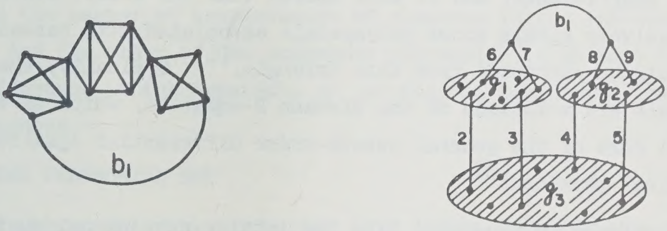

b_1
b_1
6
7
8
9
g_1
g_2
2
3
4
5
g_3

Fig. V-1. Creation of Series-pairs upon Opening b_1

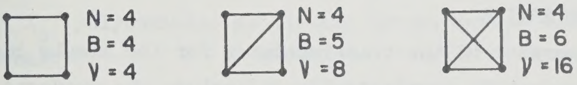

N = 4
B = 4
ν = 4
N = 4
B = 5
ν = 8
N = 4
B = 6
ν = 16

Fig. V-2. Non-degenerate Non-parallel 3-rd Order Graphs

ON THE PROPERTIES OF CERTAIN POLYNOMIALS ENCOUNTERED WITH CASCADE NETWORKS

Dr. K. A. Pullen
U. S. Army Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland 21005

INTRODUCTION

A recent paper has shown the relation of the coefficients of certain commonly encountered polynomials to the Pascal triangle of binomial coefficients, and it also showed that the coefficients of some relatively little known polynomials associated with cascade networks could be extracted from this triangle.(1) Typically, these polynomials are solutions of the Riemann P-equation, which in turn is a special form of the general second-order differential equation of mathematical physics.(2)

Some of the polynomials, like the lattice-network polynomials, are nonorthogonal in the Sturm-Liouville sense, as is evidenced by the fact the polynomial order appears as an exponent in the weighting function. All of the polynomials consist of two sets having orders differing by a half, as required by the properties of the basic differential equation which they satisfy.

TRANSFER PROPERTIES

The numerator of the transimpedance for the simple cascade network has been proven topologically to include the products of the

(1) K. A. Pullen, "On the Use of the Pascal Triangle with Network Function," First Annual Allerton Conference on Circuit and System Theory, 1963.

(2) E. T. Whittaker and G. N. Watson. Modern Analysis. Cambridge University Press, 4th ed., 1927. P. 203.

admittances of all the series, or coupling, elements of the cascade.(3) If the individual section of the cascade includes more than one path from input to output, then each path has such a product, and the total transfer function is the sum of the individual parts:

$$Y_T = \sum_{r}^{n} \prod^{j} Y_{j,r} \tag{1}$$

where the summation includes all possible combinations of elements generating transfer trees.

Typical forms for the transimpedance function, in terms of resistance and capacitance elements, are given in Table I.

Examination of the reduced Y_T terms of Table I, after elimination of common factors with Y_D, shows that the single and twin ladders, the simple tee, and the simple and the twin pi have a simple product as values, and as a result, transmission zeros cannot exist. For the twin tee and the bridged tee quotients, these expressions are quadratics, and for the lattice network, the reduced numerator is linear in "s". Clearly, the orders of the numerators terms are at most equal to the number of transmission of elements in the individual sections. And the order of the numerator polynomial is at most the maximum difference in the exponents of s for the respective transfer tree products.

THE LADDER POLYNOMIAL SET

The polynomial sets associated with ladder networks have the orthoronality property over the limits [-4,0], as can be verified by application of the Sturm-Liouville theorem to the basic differential equations:

$$s(4 + s)F_n'' + 2(s + 1)F_n' - n(n + 1)F_n = 0 \tag{2}$$

$$s(4 + s)Y_n'' + 3(s + 2)Y_n' - (n^2 - 1)Y_n = 0 . \tag{3}$$

The F_n and Y_n polynomials, as already noted, have a difference in order of one-half.

The transformation of the differential equation for the symmetric ladder network into the Chebychev form corresponds to a simple change in variable from either the tee-section polynomials or the pi-section

---

(3) See, for example, K. A. Pullen. Theory and Application of Topological and Matrix Methods. John Rider Publisher, 1962.

TABLE I - TABLE OF Y_T AND Y_D FUNCTIONS

Section	Series R Config F	Series C Config Y	Sum Y_T	Y_D
Ladder	$1/R$	sC	$C(s_o + s)$	$C^2(s_o + s)^2$
Tee	$1/R^2$	s^2C^2	$2C^2(s_o{}^2 + s^2)$	$4sC^2(s_o + s)^2/s_o$
Pi	$1/R$	sC	$C(s_o + s)$	$3C^2(s_o + s)^2$
Lattice R_1-R_2	$1/R_1{}^2$	$1/R_2{}^2$	$(1/R_{1_2}{}^2) - (1/R_2{}^2)$	$(1/R_1 + 1/R_2)^2$
Lattice R-C	$1/R^2$	s^2C^2	$C^2(s_o{}^2 - s^2)$	$C^2(s_o + s)^2$
Bridged Tee Fig 1	a/R^2	s^2C^2	$C^2(as_o{}^2 + 2ass_o + s^2)$	$sC^2(s + 2as_o)/R$
Bridged Tee Fig 2	as^2C^2	$1/R^2$	$C^2(as^2 + 2ass_o + s_o{}^2)$	$sC^2(s_o + 2as)/R$

All of the sums of the Y_T terms and the associated Y_D functions are basic elements for the polynomial sets.

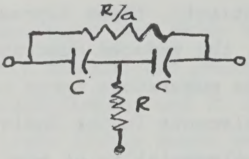


Fig 1

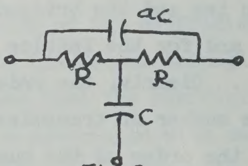


Fig 2

polynomials. Those based on the simple or "el" section are nonsymmetric, although they also are associated with Chebychev polynomials. The tee and pi sections corresponding to "Chebychev elements" are shown in Figure 3.

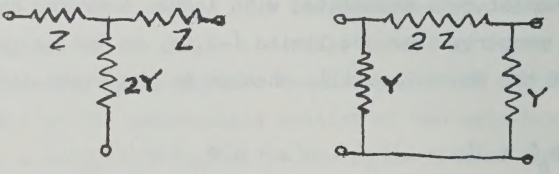


Fig 3

There is a symmetric form of Equation (2) noted by Hindin;$^{(4)}$ in this instance, the basic differential equation takes the form (after transformation to limits [-1,1]):

$$(1 - y^2)F_n'' - 3yF_n' + 4(n^2 + n)F_n = 0$$
$$(1 - y^2)Y_n'' - 3yY_n' + (4n^2 + 1)F_n = 0 \quad . \tag{4}$$

The parameters associated with the hypergeometric differential equation

(4) H. J. Hindin, "Matrix Multiplication Shortcut," Microwaves, March 1964.

in this instance are:

$$a = -2n \quad ; \quad b = 2 + 2n \quad ; \quad C = 3/2$$

$$b = 1 - j2n \; ; \; b = 1 + j2n \; ; \; c = 3/2 \quad \text{for } Y_n \quad . \qquad (5)$$

The associated polynomials are also orthorognal.

PSEUDO-BUTTERWORTH POLYNOMIALS

Standard Butterworth polynomials have the property that they have minimum slope of attenuation versus frequency over a given frequency bend; they also have the property of physical realizability. The coefficients of the high-order polynomials generally are irrational, and as a result, these polynomials are often difficult to apply. Golay(5) has developed a set of polynomials, which he calls "pseudo-Butterworth polynomials," which offer some of the properties of the Butterworth polynomials, but have integral coefficients instead of irrational ones.

The coefficients of these polynomials are more complex than those associated with the sets of orthogonal polynomials considered above; in fact, they take the form of modified hypergeometric functions:

$$C_j{}^n = (n - j)_{2j}/2^j j! + (n - j - 2)_{j2-2}/2^{j-1}(j - 1)! + \qquad (6)$$
$$(-2)^{j-1}(1/2)_{j-2} \times F\Big((2 - n - j)/2,\ (1 - n + j)/2;\ (5 - 2j)/2;\ 1\Big).$$

Clearly, this expression is rather unmanageable! It does, however, express the coefficients for the pseudo-Butterworth polynomials in closed form.

TWIN-TEE NETWORKS

The coefficients for the polynomials of the twin-tee network satisfy the recursion relation:

$$C_{r,n+1} = 2(C_{r-2,n} + 4C_{4-1,n} + C_{r,n}) - C_{r-4,n-1} - 2C_{n-2,n-1} - C_{r,n-1} \ . \qquad (7)$$

The basic coefficient triangle for this network is in Table II. The differences of order $2j + 1$ found from any diagonal in this table (r = const) are 2^{3j} or 8^j showing that the coefficients are polynomials. Unfortunately, there appears to be no simple closed-form expression for the general coefficient.

The components from which the coefficients for the twin-tee polynomials are formed can be represented in triangular form as in Table III.

(5) M. J. E. Golay, "Polynomials of Transfer Functions with Poles Only Satisfying Conditions at the Origin," IRE Transactions PGCT, CT7, No. 3 (Sep 60), p. 224.

TABLE II - COEFFICIENTS FOR TWIN-TEE POLYNOMIALS

n = 0						1	r=0				
1					1	4	1	2			
2				1	16	34	16	1	4		
3			1	36	195	328	195	36	1	6	
4		1	64	644	2240	3334	2240	644	64	1	8
5	1	100	1605	9360	25290	34904	25290	9360	1605	100	1

TABLE III - TWIN-TEE POLYNOMIALS COMPONENTS, $C_{j,r}$

n = 0						1	r=0					
1					8	-4	0	2				
2				64	-96	32	1	0	4			
3			512	-768	256	24	-8	0	0	6		
4		4096	10240	8192	-1728	-448	128	1	0	0	8	
5	32768	-81920	32768	19968	-5120	1536	48	-12	0	0	0	10

where the horizontal rows correspond to the respective values of "j," and the left-sloping diagonals sequences, such as 1, 8, 64, 512, 4096, etc., and 0, -4, -96, -768, -10240, etc., to specific values of "r." The complete coefficient takes the form:

$$A_{n,j} = \sum_{r=0}^{2j} C_{j,r}\left(n - \left[[(j+1)/2]\right] + 2j - r\right)^{(2j-r)}/(2j-r)! \qquad (8)$$

where the value of $\left[[(j+1)/2]\right]$ is the largest integer it contains. The form of the factor in the numerator is somewhat unusual, but it must take the indicated form in order to cause each coefficient series to start with the appropriate term. The complete polynomial is:

$$P_n = \sum_{j=0}^{n} A_{n,j}\, s^j\ . \qquad (9)$$

For any given value of n, the lowest order coefficient in Equation (8) is given by $n = \left[[(j+1)/2]\right]$, the value being

$$\sum_{r=0} C_{j,r}\ .$$

Higher order terms result when $n > \left[[(j+1)/2]\right]$.

THE BRIDGED-TEE NETWORK

The polynomials associated with the bridged-tee network are somewhat simpler than those for the twin-tee network, even though both are basically second order. The table of polynomial coefficients is:

TABLE IV - COEFFICIENTS FOR BRIDGED-TEE POLYNOMIALS

						j=0					
n = 0						1	2				
1					1	3	1				
2				1	8	16	8	1	4		
3			1	15	63	102	63	15	1	6	
4		1	24	164	488	702	488	164	24	1	8
5	1	35	345	1560	3750	5032	3750	1560	345	35	1

With these coefficients, the value of the term for the diagonal for which $j = 1$ is $(n^2 + 2n)$ for the bridged-tee, and $4n^2$ for the twin-tee. Higher order coefficients may be obtained by the use of the recursion relation

$$A_{n+1,j} = 2(A_{n,j-2} + A_{n,j-1} + A_{n,j}) - A_{n-1,j-4} - 4A_{n-1,j-3} - 6A_{n-1,j-2} - 4A_{n-1,j-1} - A_{n-1,j} \, . \tag{10}$$

The table can be extended as much as desired by the use of this relation.

The components of the coefficients can be expressed in tabular fashion as were those for the twin-tee. The resulting triangle is in Table V. No simple generating functions for any components $C_{n,r}$ other than that of $C_{n,o}$ have been found; the value of $C_{n,o}$ is as can be seen. The general equation for the respective coefficients is:

$$A_{n,j} = \sum_{r=0}^{2j} C_{n,r}\Big(n - \Big[[(j + 1)/2]\Big] + 2j - r\Big)^{(2j-r)}/(2j - r)! \tag{11}$$

where the same formal definitions apply as with Equation (8).

FORMATION OF SUMS

The sequences of coefficients associated with the various network polynomials can be used in the formation of a large list of summations, at least some of which apparently have not been readily available. For example, an extensive set can be generated which are associated with the "F" and the "Y" polynomials of the ladder network. Some of these are listed in Table VI.

TABLE V - BRIDGED-TEE POLYNOMIAL COMPONENTS, $C_{n,r}$

						r=0						
j = 0						1						
1					2	1	0	^{2}				
2				4	6	-10	1	0	^{4}			
3			8	28	-8	-30	6	0	0	^{6}		
4		16	72	0	-180	36	72	-15	0	0	^{8}	
5	32	208	256	-424	-480	360	92	-29	0	0	0	^{10}

TABLE VI - SUMMATION FORMULAS FROM $F_n(s)$

$$1 = F_{6n}(-1) = F_6(-1) = \sum_{r=0}^{(6n-1)} (-1)^r \binom{6n+r-1}{2r} = \sum_{r=0}^{6n} (-1)^r \binom{6n+r}{2r}$$

$$0 = F_{3n+1}(-1) = \sum_{r=0}^{3n+1} (-1)^r \binom{3n+r+1}{2r}$$

$$-1 = F_{6n+2}(-1) = F_{6n+3}(-1) = \sum_{r=0}^{6n+2} (-1)^r \binom{6n+r+2}{2r} = \sum_{mr=0}^{6n+3} (-1)^r \binom{6n+r+3}{2r}$$

$$1 = F_{4n}(-2) = F_{4n-1}(-2) = \sum_{r=0}^{4n} (-2)^r \binom{4n+r}{2r} = \sum_{r=0}^{4n-1} (-2)^r \binom{4n+r-1}{2r}$$

$$-1 = F_{4n+1}(-2) = F_{4n+2}(-2) = \sum_{r=0}^{4n+1} (-2)^r \binom{4n+r+1}{2r} = \sum_{r=0}^{4n+2} (-2)^r \binom{4n+r+2}{2r}$$

$$1 = F_{3n}(-3) = F_{3n+2}(-3) = \sum_{r=0}^{3n} (-3)^r \binom{3n+r}{2r} = \sum_{r=0}^{3n+2} (-3)^r \binom{3n+r+2}{2r}$$

$$-2 = F_{3n+1}(-3) = \sum_{r=0}^{3n+1} (-3)^r \binom{3n+r+1}{2r}$$

$$(-1)^n(2n+1) = \sum_{r=0}^{n} (-4)^r \binom{n+r}{2r}$$

where the following definitions apply:

$$F_n(s) = \sum_{r=0}^{n} \binom{n+r}{2r} s^r \tag{12}$$

$$F_{n+1}(s) = (s + 2)F_n - F_{n-1} . \tag{13}$$

A similar set of summations may be generated based on the Y_n function for the ladder network. Some of these sums are tabulated in Table VII.

TABLE VII - SUMMATION FORMULAS FOR $Y_n(s)$

$$0 = Y_{3n}(-1) = \sum_{r=0}^{3n} (-1)^r \binom{3n+r}{2r+1}$$

$$1 = Y_{6n+1}(-1) = Y_{6n+2}(-1) = \sum_{r=0}^{6n} (-1)^r \binom{6n+r+1}{2r+1} = \sum_{r=0}^{6n+1} (-1)^r \binom{6n+r+2}{2r+1}$$

$$-1 = Y_{6n+4}(-1) = Y_{6n+5}(-1) = \sum_{r=0}^{6n+3} (-1)^r \binom{6n+r+4}{2r+1} = \sum_{r=0}^{6n+4} (-1)^r \binom{6n+r+5}{2r+1}$$

$$0 = Y_{4n}(-2) = Y_{4n+2}(-2) = \sum_{r=0}^{4n-1} (-2)^r \binom{4n+r}{2r+1} = \sum_{r=0}^{4n+1} (-2)^r \binom{4n+r+2}{2r+1}$$

$$1 = Y_{4n-1}(-2) = Y_{4n+1}(-2) = \sum_{r=0}^{4n-2} (-2)^r \binom{4n+r-1}{2r+1} = \sum_{r=0}^{4n} (-2)^r \binom{4n+r+1}{2r+1}$$

$$1 = Y_{3n+1}(-3) = -Y_{3n+2}(-3) = \sum_{r=0}^{3n} (-3)^r \binom{3n+r+1}{2r+1} = \sum_{r=0}^{3n+1} (-3)^r \binom{3n+r+1}{2r+1}$$

$$0 = Y_{3n}(-3) = \sum_{r=0}^{3n-1} (-3)^r \binom{3n+r}{2r+1}$$

$$(-1)^{n-1} xn = Y_n(-4) = \sum_{r=0}^{n} (-4)^r \binom{n+r}{2r+1}$$

where $Y_n(s)$ is defined by the equation:

$$Y_n(s) = \sum_{r=0}^{n} \binom{n+r}{2r+1} s^r. \tag{14}$$

Clearly, the values of both $Y_n(0)$ and $F_n(0)$ are positive unity for all values of n.

A similar set of summations may be established for lattice polynormials, and the associated lattice polynomials which are useful in connection with the approximation of the square-root of the Laplace variable, "s." These summations include the following:(6)

---

(6) The data on the associated lattice polynomials were provided in part in a private communication from C. A. Halijak, Kansas State University.

TABLE VIII - LATTICE POLYNOMIAL SUMMATIONS

$$2^{2n-1} = G_n(1) = \sum_{r=0}^{n} \binom{2n}{2r} \; ; \; 2^{2n-1} = \sum_{r=0}^{n-1} \binom{2n}{2r+1} = Z_n(1), \; n \geq 1$$

$$1 = G_n(0) \qquad ; \qquad 2n = Z_n(0)$$

$$(-1)^n \times 2^{2n} = G_{2n}(-1) = \sum_{r=0}^{2n} \binom{4n}{2r} \; ;$$

$$(-1)^n \times (2)^{2n+1} = Z_{2n+1}(-1) = \sum_{r=0}^{2n} (-1)^r \binom{4n+2}{2r+1}$$

$$0 = G_{2n+1}(-1) = Z_{2n}(-1) = \sum_{r=0}^{2n+1} (-1)^r \binom{4n+2}{2r} = \sum_{r=0}^{2n-1} (-1)^r \binom{4n}{2r+1}$$

$$2^{2n} = \mathcal{R}_n(1) = \mathcal{T}_n(1) = \sum_{r=0}^{n} \binom{2n+1}{2r} = \sum_{r=0}^{n} \binom{2n+1}{2r+1}$$

$$1 = \mathcal{R}_n(0) \qquad ; \qquad 2n+1 = \mathcal{T}_n(0)$$

$$(-1)^n \times 2^{2n} = \mathcal{R}_{2n}(-1) = \mathcal{T}_{2n}(-1) = \sum_{r=0}^{2n} (-1)^r \binom{4n+1}{2r} = \sum_{r=0}^{2n} (-1)^r \binom{4n+1}{2r+1}$$

$$(-1)^{n+1} \times 2^{2n+1} = \mathcal{R}_{2n+1}(-1) = \mathcal{T}_{2n+1}(-1) = \sum_{r=0}^{2n+1} (-1)^r \binom{4n+1}{2r}$$

$$= \sum_{r=0}^{2n} (-1)^r \binom{4n+1}{2r+1}$$

where $$G_n(s) = \sum_{r=0}^{n} \binom{2n}{2r} s^r \qquad ; \qquad Z_n(s) = \sum_{r=0}^{n-1} \binom{2n}{2r+1} s^r \qquad (15)$$

$$\mathcal{R}_n(s) = \sum_{r=0}^{n} \binom{2n+1}{2r} s^r \qquad ; \qquad \mathcal{T}_n(s) = \sum_{r=0}^{n} \binom{2n+1}{2r+1} s^r . \qquad (16)$$

Legendre polynomials in their several symmetric and nonsymmetric forms lead to similar summations. Some of the typical ones follow:

TABLE IX - LEGENDRE POLYNOMIAL SUMMATIONS

$$1 = P_n(1) = (1/2)^{n-1} \sum_{r=0}^{[n/2]} (-1)^r \binom{2n-2r-1}{n-r}\binom{n-r}{r}$$

$$1 = P_{2n}(-1) = -P_{2n+1}(-1) = (1/2)^{2n-1} \sum_{r=0}^{n} (-1)^r \binom{4n-2r-1}{2n-r}\binom{2n-r}{r}$$

$$= (1/2)^{2n+1} \sum_{r=0}^{n} (-1)^r \binom{4n-2r+1}{2n+1-r}\binom{2n+1-r}{r}$$

$$(-1)^n = P_n(y)\big|_1 = \sum_{r=0}^{n} (-1)^r \binom{n+r}{2r}\binom{2r}{r} .$$

(This corresponds to the limit (-1) on $P_n(x)$.)

For the associated half-order polynomial $R_n(y)$,

$$n = R_n(0) = \binom{n}{1}\binom{1}{0}$$

$$(-1)^{n-1} = R_n(1) = \sum_{r=0}^{n} (-1)^r \binom{n+r}{2r+1}\binom{2r+1}{r}.$$

For the trigonometric form,

$$4^{2n+1} = P_n(0)\big|_{0,\pi} = \sum_{r=0}^{n} \binom{2n-2r-2}{n-r-1}\binom{2n+2r-1}{n+r-1}$$

where respective functions are:

$$P_n(x) = (1/2)^{n-1} \sum_{r=0}^{[n/2]} (-1)^r \binom{2n-2r-1}{n-r}\binom{n-r}{r} x^{n-2r} \quad [-1,1] \qquad (17)$$

$$P_n(y) = \sum_{r=0}^{n} \binom{n+r}{2r}\binom{2r}{r}(-y)^r \quad [0,1] \qquad (18)$$

$$R_n(y) = \sum_{r=0}^{n-1} \binom{n+r}{2r+1}\binom{2r+1}{r}(-y)^r \quad [0,1] \ . \qquad (19)$$

The last, or R form, of the polynomials is not generally known.

The Chebychev polynomials may be used as the basis of summation formulas also. The basic forms of thse polynomials are:

$$T_{2n}(x) = -2 \sum_{r=1}^{n} \left[(-r)^{r-1}(n/r)\binom{n+r-1}{2r-1}x^{2r} + \delta_{0,r}\right] \tag{20}$$

$$T_{2n+1}(x) = \sum_{r=0}^{n} (-4)^{r}\left[(2n + 1)/(2r + 1)\right]x^{2r+1} . \tag{21}$$

From these, the following summations may be obtained:

$$\begin{aligned} 0 &= \sum_{r=0}^{n} (-4)^{r-1}(n/r)\binom{n+r-1}{2r-1} \\ 1 &= \sum_{r=0}^{n} (-4)^{r}\left[(2n+1)/(2r+1)\right]\binom{n+r}{2r} . \end{aligned} \tag{22}$$

Similar expressions may be developed for the nonsymmetric Chebychev polynomials which are basic to the theory of the tee and the pi ladder networks.

CONCLUSIONS

The polynomials associated with network theory offer a fertile area for study in development of both electrical concepts and useful mathematical relations.

THE PERFECTLY COUPLED AND SHUNT-AUGMENTED T, A NEW AND VERY GENERAL TWO-PORT MODEL IN PASSIVE NETWORK REALIZATIONS

Kurt H. Haase
Air Force Cambridge Research Laboratories (OAR)
Laurence G Hanscom Field
Bedford, Massachusetts

SUMMARY

A T section in which the two series branches and part of the shunt branch are perfectly coupled, is defined and referred to as a 'perfectly coupled and shunt augmented T'. The perfectly coupled part of the T implies the arbitrary positive real frequency function $\varphi(s)$, the other part of the shunt branch, the shunt augmentation, implies another arbitrary pr frequency function $\emptyset(s)$. There are types (A),(B), and (C) of such T's. Their properties and properties tandem structures are discussed.

INTRODUCTION

Thirty-three years ago Otto Brune in his well known paper(1) (a) stated the necessary and sufficient conditions for an impedance function to be realizable as a passive network and; (b) devised a procedure for the realization. Two equivalent circuits in the Brune realization are shown in Figure 1. The impedance Z(s) is a minimum resistance function. The left one of the equivalent circuits is a model that cannot be realized since the inductances L, M, and N satisfy the equation

$$1/L + 1/M + 1/N = 0 \qquad (1)$$

and consequently one of them has to be negative. But fortunately

Equation (1) determines a perfectly coupled transformer, and therefore the circuit on the right of Figure 1 is technically realizable. However, only when a proper core material is available, when a careful winding technique is applied and when the turn ratio does not too far deviate from 1 (one), it can be expected that the technical realization behaves like the ideal model. Many successful attempts have been made in the later years to find realizations which do not imply a transformer until now when we know that such a transformless realization always exists, provided that Z(s) is a positive real function. The realization procedure according to Bott-Duffin(2) and a more recently published procedure by F. Miyata(3) start from the Brune two-port. We will in this paper present a general two-port model which under certain specializations becomes the Brune two-port. We will discuss some properties and advantages discussed in (3); however, in the most general aspect.

DEFINITION OF THE "PERFECTLY COUPLED AND SHUNT-AUGMENTED T"

Consider the T structure of a two-port as shown in Figure 2. It has the series branches with the impedances

$$U = u\cdot\varphi(s) = v\cdot(n-1)\cdot\varphi(s) \quad , \tag{2}$$

$$W = w\cdot\varphi(s) = v\cdot(\frac{1}{n} - 1)\cdot\varphi(s) \quad , \tag{2a}$$

and the shunt branch with the impedances

$$V + X = v\cdot\varphi(s) + x\cdot\emptyset(s) \quad . \tag{2b}$$

In Equations (2a, b, and c) are u, v, w, and x real constants of impedance character; the constant x in particular is positive. The notations $\varphi(s)$ and $\emptyset(s)$ are supposed to be positive real functions of the frequency variable s. Saying that the constants have impedance character allows us to consider $\varphi(s)$ and $\emptyset(s)$ as functions without dimension. We suppose that these functions are normalized in the way that in $\varphi(s) = p(s)/q(s)$ and in $\emptyset(s) = P(s)/Q(s)$ the highest power of s in the numerator and in the denominator polynomial is associated with the coefficient 1 (one). We assume that $\varphi(s)$ and $\emptyset(s)$ are allowed to be completely unrelated. We assume that the constants u, v, and w satisfy the equation

$$1/u + 1/v + 1/w = 0 \; . \tag{3}$$

Equation (3) is the same as Equation (1), using only other notations. It justifies therefore to call the T in Figure 2 a perfectly coupled one. Its shunt impedance V is augmented by the impedance X. Hence we will refer to the section shown in Figure 2 as a 'Perfectly Coupled and Shunt-Augmented T', abbreviated henceforth as pcsa-T.

THE THREE TYPES OF pcsa-T's

When in the right parts of Equations (2a, b) n is a real constant, then Equation (3) holds. This equation postulates that one of the u, v, or w has the opposite polarity of the two others. Without limitation we can agree that the one is negative. Hence, when v is positive, also n is positive, and either u or w is negative. When v is negative, also n is negative, and u and w are positive. It has some practical advantage when we introduce the positive notations $\bar{v}$ and $\bar{n}$ so that

$$\text{when } n \text{ negative } n = -\bar{n} \tag{4a}$$

$$\text{and when } v \text{ negative } v = -\bar{v} \tag{4b}$$

Out of merely formal reasons we will write x in connection with v and n, and $\bar{x}$ in connection with $\bar{v}$ and $\bar{n}$; but since the augmentation branch is always positive, $x = \bar{x}$ in the latter case. Usually henceforth are v, n, and $\bar{v}$, $\bar{n}$ considered as positive constants.

pcsa-T Type (A)	pcsa-T Type (B)	pcsa-T Type (C)
v and n positive,	v and n positive,	$v = -\bar{v}$ and $n = -\bar{n}$ negative
u positive,	u negative	($\bar{v}$ and $\bar{n}$ positive)
w negative,	w positive,	u positive
$n > 1$	$n < 1$	w positive

A pcsa-T of any type is by Equations (2a, b, and c) determined by the mutual constant v, the augmentation constant x, and the ratio constant n besides the frequency functions $\varphi(s)$ and $\phi(s)$.

THE BRUNE SECTION, A PARTICULAR pcsa-T

For the particular frequency functions $\varphi(s) = s$ and $\phi(s) = 1/s$ becomes the general pcsa-T a Brune section. Note that this section is not obtained by a frequency transformation, since $\varphi(s)$ and $\phi(s)$

are, in general, supposed to be unrelated to each other.

We will also remark that if $\varphi(s) = s$, but $\emptyset(s)$ is any positive real function $\emptyset(s) \neq 1/s$, the section can technically be realized by the use of a perfectly coupled transformer. The only difference compared with the Brune section is that the capacitive impedance $1/sC$ is replaced by the arbitrary impedance $x\,\emptyset(s)$.

THE CHAIN MATRIX OF A pcsa-T

We separate the numerator matrix elements of the ABCD matrix from a common denominator by

$$\left\| \begin{matrix} A & B \\ C & D \end{matrix} \right\| = \frac{1}{\hat{E}} \cdot \left\| \begin{matrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{matrix} \right\| \qquad (5)$$

If the two-port described by the chain matrix is a passive one then

$$AD - BC = 1 \qquad (5a)$$

$$\hat{A}\hat{D} - \hat{C}\hat{D} = \hat{E}^2 \qquad (5b)$$

hold.

From the well known chain matrix of a T section we derive:

$$\hat{A} = v\,n\,\varphi(s) + x\,\emptyset(s)\ , \qquad (6a)$$

$$\hat{B} = v\,\frac{(n-1)^2}{n}\,\varphi(s)\,\emptyset(s)\ , \qquad (6b)$$

$$\hat{C} = 1 \qquad (6c)$$

$$\hat{D} = \frac{v}{n}\,\varphi(s) + x\,\emptyset(s)\ , \qquad (6d)$$

$$\hat{E} = v\,\varphi(s) + x\,\emptyset(s)\ . \qquad (6e)$$

Equations (6a, ..., e) give the matrix elements of a pcsa-T of type (A) or (B); for type (C) the matrix elements are obtained by replacing $n = -\bar{n}$ and $v = -\bar{v}$.

TRANSPOSITION OF A PROPER IMPEDANCE BETWEEN THE PORTS OF A pcsa-T

Consider a pcsa-T of types (A) or (B) as shown in Figure 3. The T on the left side of the figure, which besides the frequency functions $\varphi(s)$ and $\emptyset(s)$ is determined by the constants v, n, and x, has a shunt impedance $X_s = x_s\,\emptyset(s)$ across its in-port. The T on the right side of the figure which implies the same frequency functions and is determined by the constants v', n', and x' has a shunt impedance $X'_s = x'_s\emptyset(s)$

across its out-port. The constants x_s and x'_s are supposed to be positive. By simple matrix algebra we find the following elements of the chain matrix of the left side T in Figure 3:

$$\hat{A} = n\varphi(s) + \frac{x}{v}\emptyset(s) \tag{7a}$$

$$\hat{B} = x\frac{(n-1)^2}{n}\varphi(s)\cdot\emptyset(s) \tag{7b}$$

$$\hat{C} = \frac{n}{x_s}\cdot\frac{\varphi(s)}{\emptyset(s)} + \frac{x + x_s}{vx_s} \tag{7c}$$

$$\hat{D} = \frac{1}{n}\left[\frac{x}{x_s}\cdot(n-1)^2 + 1\right]\cdot\varphi(s) + \frac{x}{v}\emptyset(s) , \tag{7d}$$

$$\hat{E} = \varphi(s) + \frac{x}{v}\emptyset(s) . \tag{7e}$$

The elements of the matrix of the T on the right side in Figure 3 are:

$$\hat{A}' = \left[n' + \frac{x'}{x'_s}\;\frac{(n'-1)^2}{n'}\right]\cdot\varphi(s) + \frac{x'}{v'}\emptyset(s) , \tag{8a}$$

$$\hat{B}' = x'.\frac{(n'-1)^2}{n'}.\varphi(s).\emptyset(s) , \tag{8b}$$

$$\hat{C}' = \frac{1}{x'_s n'}\frac{\varphi(s)}{\emptyset(s)} + \frac{x' + x'_s}{v'x'_s} , \tag{8c}$$

$$\hat{D}' = \frac{\varphi(s)}{n'} + \frac{x'}{v'}.\emptyset(s) , \tag{8d}$$

$$\hat{E}' = \varphi(s) + \frac{x'}{v'}\emptyset(s) . \tag{8e}$$

The two circuits in Figure 3 are equivalent if they have the same chain matrix. We obtain the conditions for this equivalence when we compare the respective elements $\hat{A}, \ldots, \hat{E}$ with the elements $\hat{A}', \ldots, \hat{E}'$. In order to be able to make this comparison we divided all elements in Equations (7a, ..., e) by v and all elements in Equations (8a, ..., e) by v'. The results of the comparisons are:

$$n' = n\cdot\frac{x_s}{x_s + x(n-1)^2} , \tag{9a}$$

$$x'_s = \frac{x_s}{nn'} , \tag{10a}$$

$$\frac{1}{x'} = \frac{1}{x} + \frac{1}{x_s} - \frac{1}{x'_s} , \tag{11a}$$

$$v' = v\frac{x'}{x} . \tag{12a}$$

These equations can be reversed to

$$n = n' + \frac{(n' - 1)^2}{n'} \cdot \frac{x'}{x'_s} \quad , \tag{9b}$$

$$x_s = x'_s nn' \quad , \tag{10b}$$

$$\frac{1}{x} = \frac{1}{x'} + \frac{1}{x'_s} - \frac{1}{x_s} \quad , \tag{11b}$$

$$v = v' \frac{x}{x'} \quad . \tag{12b}$$

The sections shown in Figure 3 may also both be of type (C). In this event the same Equations (9a, ..., 12a) and (9b, ..., 12b) are true when we replace v, n, and x by $-\bar{v}$, $-\bar{n}$, and $\bar{x}$.

The equivalence can also be interpreted as a transposition of the port impedance from one port to the other. Hence we can state:

THEOREM 1

If a pcsa-T has a port impedance that implies the same frequency function as its shunt augmentation, then this impedance can be transposed to the other port. The transposed impedance and the pcsa-T in the way of the transposition changes its constants from unprimed to primed values and vice versa according to Equations (9a, ..., 12a) and (9b, ..., 12b).

Note that according to Equations (12a) and (12b) the ratios v/x and v'/x' are the same. The most important effect of the transposition -- as we will see later -- is that in maximum, if we transpose the entire port impedance, the ratio constant can be changed between n and n'.

A TANDEM OF TWO pcsa-T's, THE ONE OF TYPES (A) OR (B), THE OTHER OF TYPE (C).

Assume a tandem of two pcsa-T's, as shown left in Figure 4, where T_a is followed by T_b. Let T_a be of type (A) or (B) so that $n_a = n_0$ is positive. Let T_b be of type (C) so that its ratio constant is negative and assume that $\bar{n}_b = -n_b = 1/n_0$. Assume further that the ratios $v_a/x_a = \bar{v}_b/\bar{x}_b = K$. Assume another tandem of two pcsa-T's as shown right in Figure 4, where T_c is followed by T_d. Let T_c be of type (C) so that its ratio constant $n_c = -\bar{n}_c = -n_0$ is negative. Let

T_d be of type (B) if T_a is of type (A), or of type (B) if T_a is of type (B), so that $n_d = 1/n_0$. Assume further that $\bar{v}_c/\bar{x}_c = v_d/x_d = K$. Then any of the T's determines the three others according to

$$n_0 = n_a = 1/\bar{n}_b = \bar{n}_c = 1/n_d \quad , \tag{13}$$

$$v_a/x_a = \bar{v}_b/\bar{x}_b = \bar{v}_c/\bar{x}_c = v_d/x_d = K \quad , \tag{14}$$

Let us introduce

$$v_a/\bar{v}_b = x_a/\bar{x}_b = k_a \quad , \tag{15}$$

$$v_d/\bar{v}_c = x_d/\bar{x}_c = k_b \quad . \tag{16}$$

The constants K, k_a, and k_b are positive.

The elements of the chain matrix of the left tandem in Figure 4 are

$$\hat{A}_{ab} = \phi^2(s) + K^2 \cdot \varphi^2(s) + + \frac{K}{n_0} \cdot \left[(n_0^2 + 1) + k_a \cdot (n_0 - 1)^2 \right] \cdot \varphi(s) \cdot \phi(s) \ , \tag{17a}$$

$$\hat{B}_{ab} = \frac{v_a}{k_a} \cdot \left\{ K \left[(n_0 + 1)^2 + k_a \cdot (n_0 - 1)^2 \right] \cdot \varphi(s) + + \frac{1}{n_0} \cdot \left[(n_0 + 1)^2 + k_a \cdot (n_0 - 1)^2 \right] \cdot \phi(s) \right\} \cdot \varphi(s) \cdot \phi(s), \tag{17b}$$

$$\hat{C}_{ab} = \frac{K}{n_0 v_a} \cdot (k_a + 1) \cdot \left[K \cdot \varphi(s) + n_0 \cdot \phi(s) \right] , \tag{17c}$$

$$\hat{D}_{ab} = \phi^2(s) + K^2 \cdot \varphi^2(s) + + \frac{K}{n_0} \cdot \left[(n_0^2 + 1) + \frac{1}{k_a} \cdot (n_0 + 1)^2 \right] \cdot \varphi(s) \cdot \phi(s) \ , \tag{17d}$$

$$\hat{E}_{ab} = \phi^2(s) - K^2 \cdot \varphi^2(s) \ . \tag{17e}$$

The elements of the chain matrix of the right tandem in Figure 4 are:

$$\hat{A}_{cd} = \phi^2(s) + K^2 \cdot \varphi^2(s) + + \frac{K}{n_0} \cdot \left[(n_0^2 + 1) + \frac{1}{k_b} \cdot (n_0 + 1)^2 \right] \cdot \varphi(s) \cdot \phi(s) \ , \tag{18a}$$

$$\hat{B}_{cd} = v_b \cdot \left\{ K \cdot \left[(n_0 - 1)^2 + \frac{1}{k_b} \cdot (n_0 + 1)^2 \right] \cdot \varphi(s) + + \frac{1}{n_0} \cdot \left[(n_0 - 1)^2 + \frac{1}{k_b} \cdot (n_0 + 1)^2 \right] \cdot \phi(s) \right\} \cdot \varphi(s) \cdot \phi(s), \tag{18b}$$

$$\hat{C}_{cd} = \frac{K}{n_0 v_b} \cdot (k_b + 1) \cdot \left[K \cdot \varphi(s) + n_0 \cdot \phi(s) \right] , \tag{18c}$$

$$\hat{D}_{cd} = \emptyset^2(s) + K^2 \cdot \varphi^2(s) +$$

$$+ \frac{K}{n_0} \cdot \left[(n_0^2 + 1) + k_b \cdot (n_0 - 1)^2 \right] \cdot \varphi(s) \cdot \emptyset(s) \; , \tag{18d}$$

$$\hat{E}_{cd} = \emptyset^2(s) - K^2 \cdot \varphi^2(s) \; . \tag{18e}$$

By comparing the elements in Equations (17a,..., e) with those in Equations (18a,..., e) we find that they become identical and thus both tandems in Figure 4 are equivalent if

$$k_a k_b = \left[\frac{n_0 + 1}{n_0 - 1} \right]^2 \; , \tag{19}$$

and $v_a / v_d = x_a / x_d = \dfrac{k_a + 1}{k_b + 1}$. (20)

Equations (13),..., (16), (19) and (20) are sufficient to determine one equivalent in Figure 4 from the other, and we can state:

THEOREM 2

A tandem of two pcsa-T's of type (A) or (B) followed by type (C) is equivalent with another tandem of two pcsa-T's of type (C) followed by type (B) or (A) if in each tandem the ratio constants of the T's are negative reciprocal, if the ratios $|v/x|$ of the T's in both tandems are the same, and if the ratio constants of the T's of type (C) are reciprocal.

A LATTICE EQUIVALENCE

Assume a lattice two-port as shown in Figure 5. This lattice has a pair of impedance branches

$$X_1 = x_1 \, \varphi(s) \; , \tag{21a}$$

$$X_4 = x_4 \, \varphi(s) \; , \tag{21b}$$

and another pair of impedance branches

$$X_2 = x_2 \, \emptyset(s) \; , \tag{22a}$$

$$X_3 = x_3 \, \emptyset(s) \; . \tag{22b}$$

In Equations (21a, b) are x_1 and x_4 positive constants of impedance character and $\varphi(s)$ is the same frequency function as implied in the coupled part of the pcsa-T. In Equations (22a, b) are x_2 and x_3 positive constants of impedance character and $\emptyset(s)$ is the same frequency function as implied in the shunt augmentation. Such a lattice has a chain matrix with the following elements:

$$\hat{A}_x = \phi^2(s) + \frac{x_1x_4}{x_2x_3}\varphi^2(s) + \frac{x_1x_2 + x_3x_4}{x_2x_3} \cdot \varphi(s) \cdot \phi(s) \quad , \tag{23a}$$

$$\hat{B}_x = \left[\frac{x_1x_4}{x_2x_3}(x_2 + x_3) \cdot \varphi(s) + (x_1 + x_4) \cdot \phi(s) \right] \cdot \varphi(s) \cdot \phi(s) \quad , \tag{23b}$$

$$\hat{C}_x = \frac{x_1 + x_4}{x_2x_3} \cdot \varphi(s) + \frac{x_2 + x_3}{x_2x_3} \cdot \phi(s) \quad , \tag{23c}$$

$$\hat{D}_x = \phi^2(s) + \frac{x_1x_4}{x_2x_3}\varphi^2(s) + \frac{x_1x_3 + x_2x_4}{x_2x_3} \cdot \varphi(s) \cdot \phi(s) \quad , \tag{23d}$$

$$\hat{E}_x = \phi^2(s) - \frac{x_1x_4}{x_2x_3} \cdot \varphi^2(s) \quad . \tag{23e}$$

By comparing the matrix elements in Equations (23a, . . . , e) with those in Equations (17a, . . . , e) or in Equations (18a, . . . , e) it can be shown that the lattice in Figure 5 is equivalent with the tandems in Figure 4. It is therefore possible to compute the constants of the lattice when the constants of the equivalent tandem are known and vice versa. For this purpose we introduce

$$S = K \cdot (x_2 + x_3) = \frac{v_a}{k_a}\left[(n_0 + 1)^2 + k_a(n_0 - 1)^2 \right] , \tag{24a}$$

$$= \frac{v_d}{k_b}\left[(n_0 + 1)^2 + k_b(n_0 - 1)^2 \right] \tag{24b}$$

and $P = \dfrac{x_2 + x_3}{x_2x_3} = \dfrac{k_a + 1}{v_a} K$, (25a)

$$= \frac{k_b + 1}{v_d} K \; . \tag{25b}$$

In the lattice

$$K = +\sqrt{\frac{x_1x_4}{x_2x_3}} \quad . \tag{26}$$

Equations (24a, b) hold when a pcsa-T of type (A) or (B) is followed by one of type (C). Equations (25a, b) hold for the inverse sequence. For the tandems k_a and k_b are known by Equations (15) and (16), and K is known by Equation (14). The ratio n_0 is known by Equation (13).

When the tandems are equivalent with the lattice and the tandem is known we find the constants of the lattice as

$$x_{1,4} = \frac{S}{2n_0}\left[\, 1 \pm \sqrt{1 - \frac{4Kn_0^2}{SP}}\,\right], \tag{27}$$

$$x_{2,3} = \frac{S}{2K}\left[\, 1 \pm \sqrt{1 - \frac{4K}{SP}}\,\right]. \tag{28}$$

In Equation (27) associate x_1 with the + sign.

In Equation (28) associate x_2 with the + sign if $k_a,\ k_b > \left|\dfrac{n_0 + 1}{n_0 - 1}\right|$;

otherwise, associate it with the minus sign.

When the tandem should be derived from the lattice

$$n_0 = \frac{x_2 + x_3}{x_1 + x_4}\, K\,, \tag{29}$$

$$k_a = \frac{1}{(n_0 - 1)^2} \cdot \left[\, \frac{x_1x_2 + x_3x_4}{x_2x_3}\ \frac{n_0}{K} - (n_0^2 + 1)\,\right], \tag{30a}$$

$$\frac{1}{k_b} = \frac{1}{(n_0 + 1)^2} \cdot \left[\, \frac{x_1x_2 + x_3x_4}{x_2x_3}\ \frac{n_0}{K} - (n_0^2 + 1)\,\right]. \tag{30b}$$

We are able to state:

THEOREM 3

The tandems which are equivalent according to theorem 2 are also equivalent with a lattice two-port in which the one pair of impedance branches implies the frequency function $\varphi(s)$, the other the function $\emptyset(s)$ as shown in Figure 4.

NEGATIVE RECIPROCAL RATIO CONSTANTS IN A TANDEM OBTAINED BY A TRANSPOSITION OF A PORT IMPEDANCE

Assume that in the tandem left in Figure 6 the first pcsa-T (a) is of type (A) or (B) and the second (b) of type (C). Assume further that both T's have the same ratio $|v/x|$, but that $n_a \neq 1/\bar{n}_b$. Assume that at the in-port an impedance $X_s = x_s \cdot \emptyset(s)$ is available. Then this impedance can be transposed to the out-port where it appears as X_s' as shown in the right in Figure 6. The transposition changes the unprimed constants of the tandem to primed ones according to Equations (9a, . . . , 12a). If a proper x_s is available, we may achieve that after the transposition the pcsa-T's become such that $n_a' = 1/\bar{n}_b'$

and then the tandem has a lattice equivalent with the load X'_s. It can be shown that for this purpose x_s has to be

$$x_s = \frac{1}{n_a \bar{n}_b - 1} \cdot \left[x_a (n_a - 1)^2 + \bar{x}_b n_a^2 (\bar{n}_b + 1)^2 \right] \tag{31a}$$

Since x_s must be positive, the aim can only be achieved if before the transposition $n_a \bar{n}_b > 1$. If this is the case, Equation (31a) determines the amount of x_s. If the sequence of the T's is inverse so that a type (C) is followed by a type (A) or (B), then

$$x_s = \frac{1}{\bar{n}_a n_b - 1} \cdot \left[\bar{x}_a (\bar{n}_a + 1)^2 + x_b \bar{n}_a^2 (n_b - 1)^2 \right] \tag{31b}$$

Also in this event $\bar{n}_a n_b > 1$ is necessary.

If inversely an impedance X'_s is available at the out-port and if $n'_a \neq 1/\bar{n}'_b$, the ratios $n_a = /1\bar{n}_b$ may be obtained by transposing X'_s to the in-port, where it appears as X_s. In order to obtain negative reciprocal $n_a = -1/n_b$ it is necessary that

$$x'_s = \frac{1}{n'_a \bar{n}'_b (1 - n'_a \bar{n}'_b)} \left[x'_a (n'_a - 1)^2 + \bar{x}'_b n'^2_a (\bar{n}'_b + 1)^2 \right] \tag{32a}$$

and in the event of the inverse sequence of the T's, that

$$x'_s = \frac{1}{\bar{n}'_a n'_b (1 - \bar{n}'_a n'_b)} \left[\bar{x}'_a (\bar{n}'_a + 1)^2 + x'_b \bar{n}'^2_a (n'_b - 1)^2 \right] . \tag{32b}$$

Hence, in the event of a transposition from the out-to the in-port it is necessary that before the transposition the products $n_a \bar{n}'_b$ and $\bar{n}'_a n'_b$, respectively, are smaller than 1 (one). In all these cases can the constant of the port impedance be smaller than the values computed from Equations (31a, b) and (32a, b), since we split off x_s and x'_s in parallel. In these events x_s and x'_s would perform the transposition and the remaining part of the port impedance would be left at the respective port.

ON THE REALIZATION OF A pcsa-T

A pcsa-T can be realized by using a perfectly coupled transformer if the frequency function $\varphi(s) = s$. In all other cases the pcsa-T can only be considered as a non-realizable model. The tandems, however, can be realized by the lattice equivalence. The realization of a pcsa-T by the Bott-Duffin method$^{(2)}$ is possible if $\emptyset(s) = 1/\varphi(s)$ and if a frequency transformation

$$\sigma = \varphi(s) \tag{33}$$

exists so that $Z(\sigma)$ is positive real and minimum resistance. Such a realization is shown in Figure 7. The Richards function used in the realization is

$$R(s) = \frac{s_0 Z(s) - Z(s_0)\varphi(s)}{s_0 Z(s_0) - \varphi(s) Z(s)} \tag{34a}$$

and its inversion

$$Z(s) = Z(s_0)\,\frac{\varphi(s) + s_0 R(s)}{s_0 + \varphi(s) R(s)} \ . \tag{34b}$$

If $Z(s) = p(s)/q(s)$ and $\varphi(s) = \alpha(s)/\beta(s)$, then s_0 is the negative solution of the polynomial equation

$$\frac{p(s)\beta(s) - u\cdot q(s)\alpha(s)}{v\alpha^2(s) + x\beta^2(s)} = 0 \quad , \tag{35}$$

where u, v, and x are the impedance constants of the pcsa-T.

THE DUAL OF A pcsa-T

The dual of a pcsa-T is a perfectly coupled and series augmented Pi. The admittance V* that is one part of its series branch admittance is perfectly coupled with the shunt admittances U* at the output and W* at the input side. These admittances imply the frequency function $\varphi(s)$. The other part of the series branch is the augmentation X* that is also an admittance and implies the function $\emptyset(s)$. All previous discussions hold for this Pi to-port when the well known rules of duality are correctly applied.

CONCLUSION

Besides the fact that the application of the pcsa-T in the realization procedure of impedances is a future problem, the pcsa-T can be considered as a very general and probably useful model of a two-port.

REFERENCES

1. Otto Brune, "Synthesis of a Finite Two-Terminal Network Whose Driving-Point Impedance is a Prescribed Function of Frequency", J. Math. and Phys., Vol. 10 (Aug, 1931), pp 191 - 236.

2. R. Bott and R. J. Duffin, "Impedance Synthesis Without Use of Transformers", J. Appl. Phys. Vol. 20 (1949), p. 816.

3. Fusachika Miyata, "Synthesis of a One-Port Utilizing a Lattice Construction", IEEE Transactions on Circuit Theory, Vol. CT-10 (June, 1963), pp 227-234.

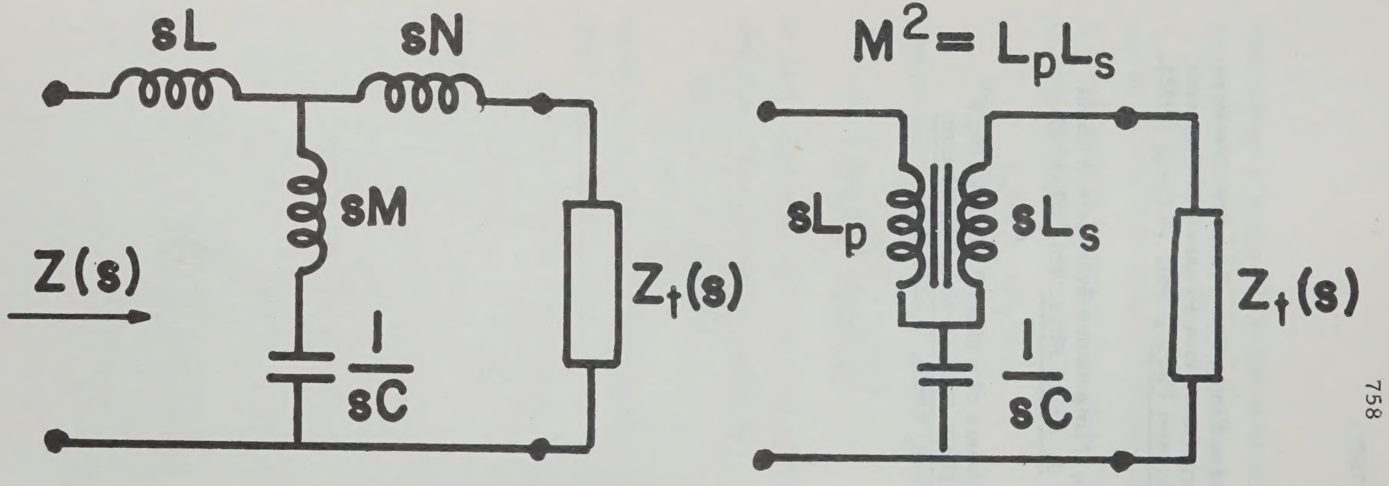

sL
sN
sM
Z(s)
1/sC
Z_t(s)
M^2= L_pL_s
sL_p
sL_s
1/sC
Z_t(s)

FIGURE 1

BRUNE REALIZATION

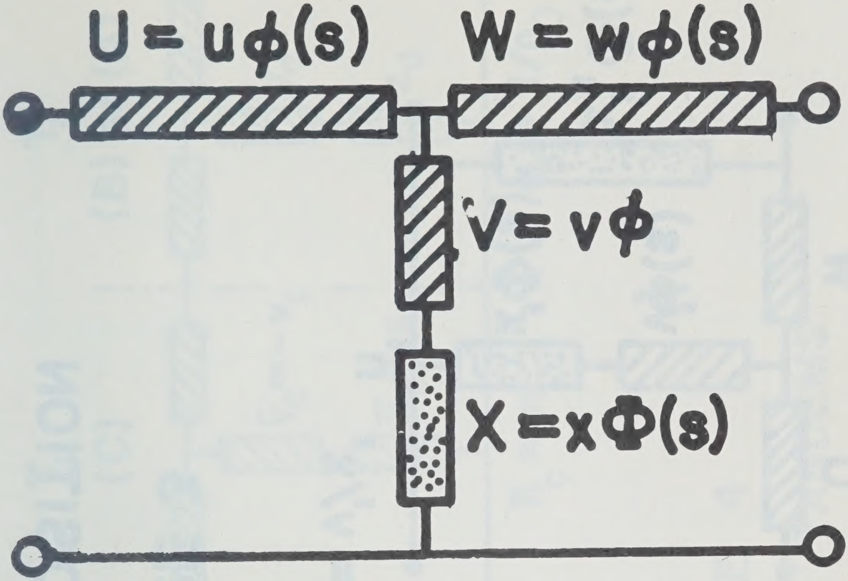

U = uφ(s)
W = wφ(s)
V = vφ
X = xΦ(s)

FIGURE 2

PERFECTLY COUPLED
AND
SHUNT-AUGMENTED T
(pcsa-T)

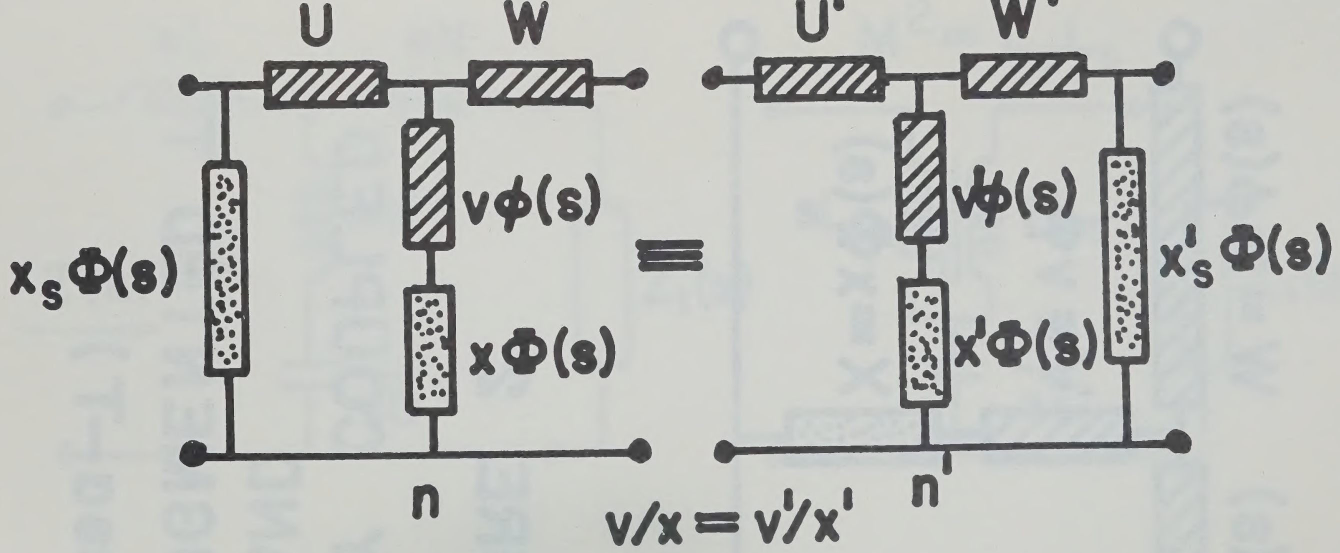

U
W
U'
W'
$x_s\Phi(s)$
$v\phi(s)$
$x\Phi(s)$
≡
$v'\phi(s)$
$x'\Phi(s)$
$x'_s\Phi(s)$
n
n'
$v/x = v'/x'$

FIGURE 3

TRANSPOSITION
OF
PORT IMPEDANCES

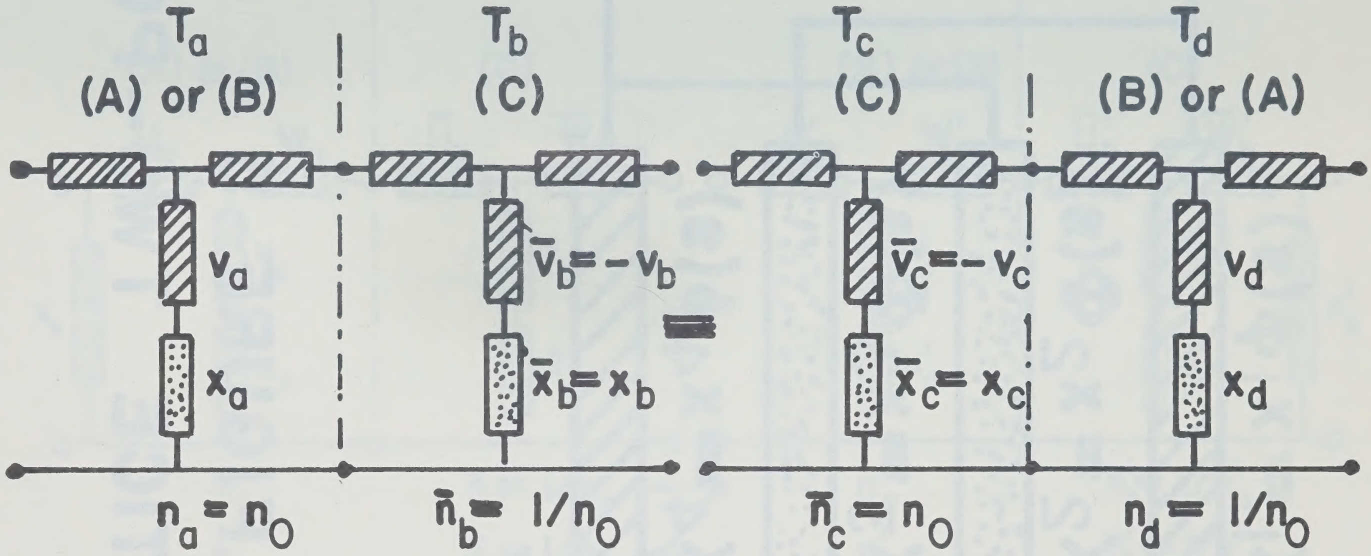

T_a
(A) or (B)
v_a
x_a
$n_a = n_O$
T_b
(C)
$\bar{v}_b = -v_b$
$\bar{x}_b = x_b$
$\bar{n}_b = 1/n_O$
$\equiv$
T_c
(C)
$\bar{v}_c = -v_c$
$\bar{x}_c = x_c$
$\bar{n}_c = n_O$
T_d
(B) or (A)
v_d
x_d
$n_d = 1/n_O$

FIGURE 4

EQUIVALENT TANDEMS

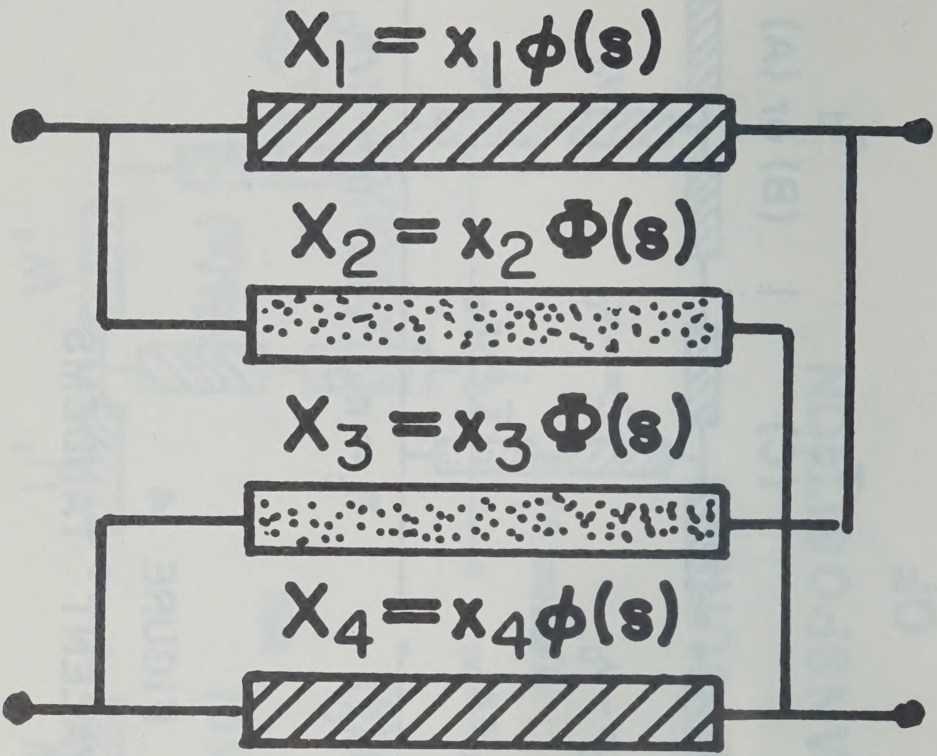

$X_1 = x_1 \phi(s)$
$X_2 = x_2 \Phi(s)$
$X_3 = x_3 \Phi(s)$
$X_4 = x_4 \phi(s)$

FIGURE 5
LATTICE TWO-PORT

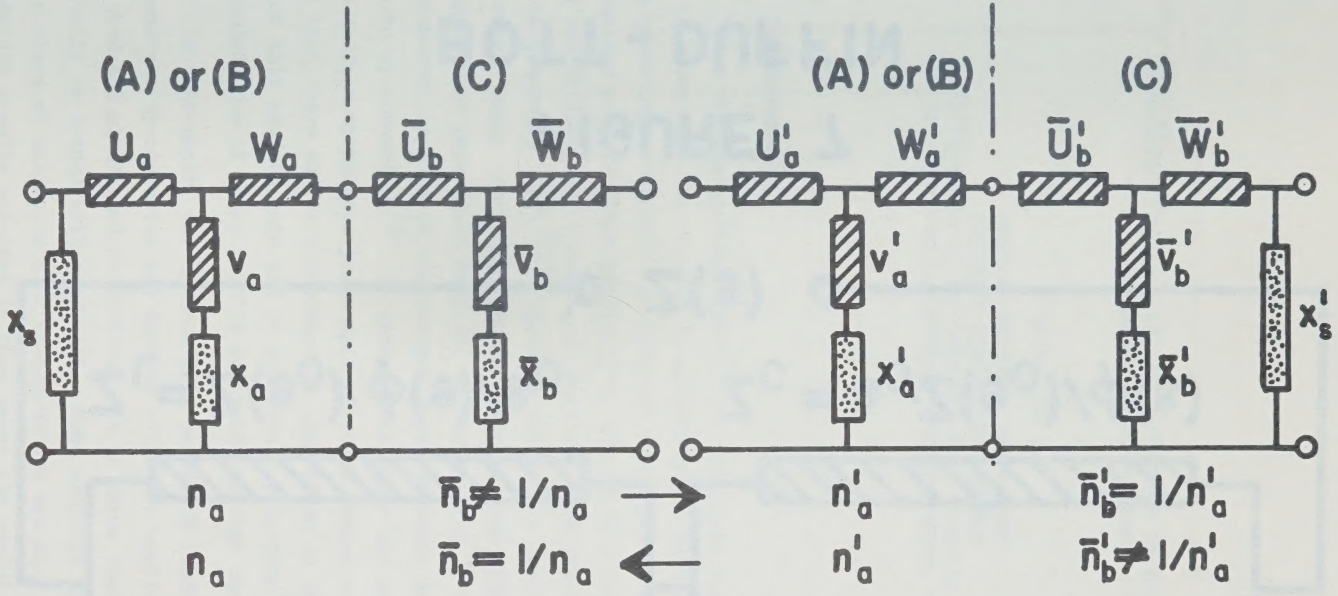

(A) or (B)
(C)
(A) or (B)
(C)
U_a
W_a
$\bar{U}_b$
$\bar{W}_b$
U'_a
W'_a
$\bar{U}'_b$
$\bar{W}'_b$
x_s
v_a
x_a
$\bar{v}_b$
$\bar{x}_b$
v'_a
x'_a
$\bar{v}'_b$
$\bar{x}'_b$
x'_s
n_a
$\bar{n}_b \neq 1/n_a$
n'_a
$\bar{n}'_b = 1/n'_a$
n_a
$\bar{n}_b = 1/n_a$
n'_a
$\bar{n}'_b \neq 1/n'_a$

FIGURE 6

ACHIEVEMENT OF A TANDEM WITH NEGATIVE RECIPROCAL RATIO CONSTANTS

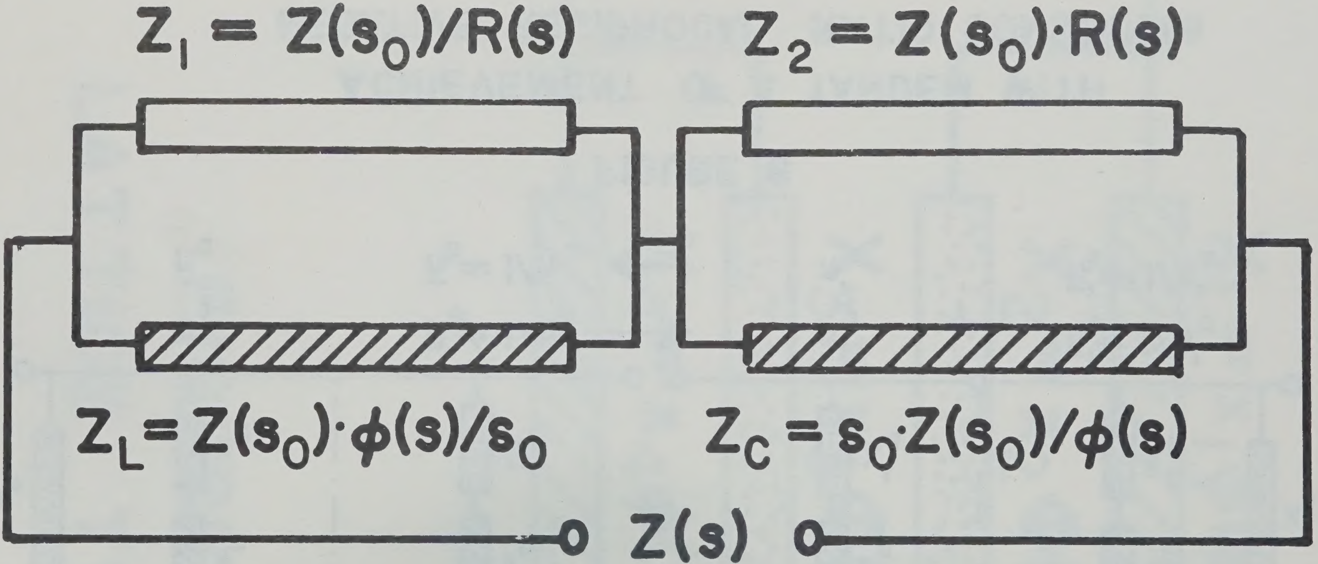

$Z_1 = Z(s_0)/R(s)$
$Z_2 = Z(s_0)\cdot R(s)$
$Z_L = Z(s_0)\cdot\phi(s)/s_0$
$Z_C = s_0\cdot Z(s_0)/\phi(s)$
$Z(s)$

FIGURE 7
BOTT - DUFFIN
REALIZATION

FUNDAMENTAL GAIN-BANDWIDTH LIMITATIONS OF NON-INVERTING PARAMETRIC UP-CONVERTERS

By
Walter H. Ku

APPLIED RESEARCH LABORATORY
SYLVANIA ELECTRONIC SYSTEMS
A Division of Sylvania Electric Products Inc.
40 SYLVAN ROAD, WALTHAM, MASSACHUSETTS, 02154

1. INTRODUCTION

From the Manley-Rowe theory(1) of nonlinear reactance amplifiers, it is known that the maximum transducer power gain obtainable with a non-inverting parametric up-converter operating in an ideal three-frequency mode is given precisely by the ratio ω_u/ω_s, where ω_s and ω_u are respectively the signal and upper-sideband frequencies. Moreover, it is generally known(2,3) that, by using more complicated equalizer networks, wider bandwidths can be achieved than with the conventional single-tuned parametric devices. Recently, Kuh(4) has published some fundamental results on the maximum gain-bandwidth product of a parametric up-converter.(5) He showed that under the matched condition, the maximum transducer power gain given by Manley and Rowe is attained and that there exists a fundamental limitation on the bandwidth over which this maximum gain can be achieved. Helgesson(6) has further extended some of these results under the matched condition.

In this paper, the optimum gain-bandwidth limitations for the parametric up-converter are derived which are applicable under both matched and mismatched conditions. It is shown that the mismatch can actually be traded with bandwidth. Some explicit results for the design of a class of maximally flat up-converters are also presented. The gain-bandwidth results obtained in this paper are useful in the synthesis of broadband parametric up-converters.

Within the scope of the small-signal analysis, the pumped varactor appears to the small signals as a linear time-variable capacitor. We assume that all the sideband frequencies are short-circuited except the signal and upper-sideband frequencies. For simplicity, we further assume a dissipationless model for the pumped varactor.

2. BASIC RELATIONS FOR PARAMETRIC UP-CONVERTERS

Figure 1 depicts a varactor placed across an otherwise arbitrary passive lossless three-port in which the signal frequency, ω_s, and the upper-sideband frequency, $\omega_u = \omega_o + \omega_s$, ($\omega_o$ = pump frequency), can exist. Let C_o and C_1 denote the first two coefficients of an arbitrary linear variable capacitance, $C(t) = C_o + 2C_1 \cos \omega_o t + \cdots$. The constant capacitance contribution, C_o, is shown explicitly shunting the equalizer. The diode impedance at the signal frequency $z_D(j\omega_s)$ is related to $Z_3(j\omega_u)$, the impedance looking into the equalizer at the diode port at the upper-sideband frequency, by the following equation

$$z_D(j\omega_s) = \frac{1}{\omega_s \omega_u C_1^2 Z_3(j\omega_u)} \quad . \tag{1}$$

Since $Z_3(j\omega_u) = Z_3(j\omega_o + j\omega_s)$ is a positive function and

$$\mathrm{Re}\; z_D(j\omega_s) = \frac{1}{\omega_s(\omega_o + \omega_s)C_1^2} \mathrm{Re}\; Y_3(j\omega_o + j\omega_s) \quad , \tag{2}$$

it is clear that the varactor diode impedance exhibits a positive real part within the frequency band of interest.

A scattering formalism using complex and frequency dependent normalization$^{(7)}$ is employed to describe the 3-port N (see Figure 1). The normalizing impedances for the three ports are selected as R_g, R_L and z_D respectively. If the 3-port N is lossless and all the normalizing impedances have positive real parts over the frequency band of interest, then the "normalized" 3 x 3 scattering matrix S is unitary over this band; i.e., $1 - \bar{S}' S = 0$.

From Figure 1, the transducer power gain from port 1 to port 2 can be expressed precisely by the square of the magnitude of the transfer coefficient s_{21}, i.e.,

$$G_{12}(\omega_s) = \frac{R_L\left|I_2(j\omega_s)\right|^2}{\frac{\left|E_g(\omega_s)\right|^2}{4R_g}} = \left|s_{21}(j\omega_s)\right|^2 \tag{3}$$

Equation (3) gives the transducer power gain when the output power is taken at the signal frequency. From the unitary condition, $\left|s_{21}(j\omega_s)\right|^2 \leq 1$; clearly, this case is of no practical interest. When the output power is taken at the upper-sideband frequency, the converter transducer power gain is given by

$$G_t(\omega_s) = \frac{\omega_u}{\omega_s}\left|s_{31}(j\omega_s)\right|^2 \left|s_{23}(j\omega_u)\right|^2 , \tag{4}$$

where ω_u/ω_s is the Manley-Rowe factor for the up-converter, and $\left|s_{23}(j\omega_u)\right|^2$ accounts for the power transfer from the varactor port to the load at the upper-sideband frequency.

In order to provide the needed isolation, a 3-port circulator is used to connect the generator, the varactor, and the load.* A lossless reciprocal 2-port equalizer is placed between the circulator and the varactor and the nonreciprocal up-converter is shown schematically in Figure 2. Now, all the upper-sideband power reflected from the varactor is absorbed by the load, and the steady-state converter transducer power gain, Equation (4), simplifies to

$$G_t(\omega_s) = \frac{\omega_u}{\omega_s}\left|s_{31}(j\omega_s)\right|^2 = \frac{\omega_u}{\omega_s}\left(1 - \left|s_{33}(j\omega_s)\right|^2\right) . \tag{5}$$

The scattering coefficient at the diode port, s_{33}, can be expressed in terms of Z_3, which represents the impedance seen looking into the diode port when ports 1 and 2 are terminated in their respective normalizing impedances, and $z_3 = z_D$, which is the normalizing impedance at the diode port. We have

$$s_{33} = \frac{Z_3 - \overline{z}_3}{Z_3 + z_3} = \frac{Z_3 - \overline{z}_D}{Z_3 + z_D} , \tag{6}$$

---

*Isolation may also be provided by using balanced circuits with four identical varactors in which case the bandwidth achievable will be twice of that derived in this paper (see Reference 6).

$$s_{33}(j\omega_s) = \frac{Z_3(j\omega_s) - \bar{z}_D(j\omega_s)}{Z_3(j\omega_s) + z_D(j\omega_s)} = \frac{\omega_s\omega_u C_1^2 - Y_3(j\omega_s)\bar{Y}_3(j\omega_u)}{\omega_s\omega_u C_1^2 + Y_3(j\omega_s)Y_3(j\omega_u)} \quad . \qquad (7)$$

The expression (7) demonstrates clearly the quadratic nature of the scattering coefficient at the diode port; i.e., s_{33} is quadratic in Y_3. This property is due to the interaction between the signal and the sideband circuit through the variable capacitor and is present for the non-inverting as well as the inverting parametric device.(8) In fact, the broadbanding problem treated in this paper is somewhat analogous to that of broadband equalization of two partially constrained passive networks.(9,10)

Substituting Equation (7) into Equation (5), we have

$$G_t(\omega_s) = \frac{\omega_u}{\omega_s}\left(1 - \left|\frac{\omega_s\omega_u C_1^2 - Y_3(j\omega_s)\bar{Y}_3(j\omega_u)}{\omega_s\omega_u C_1^2 + Y_3(j\omega_s)Y_3(j\omega_u)}\right|^2\right) . \qquad (8)$$

It can be concluded that the ultimate power gain for the parametric converter is completely delimited by the equalizer impedance Z_3 facing the varactor at both the signal and upper-sideband frequencies.

3. GAIN-BANDWIDTH LIMITATIONS

The gain-bandwidth performance of the parametric up-converter shown in Figure 2 depends on the scattering coefficient s_{33} seen looking back from the varactor towards the reactive equalizing network. We shall proceed by using Equations (5) and (8) to optimize the gain in terms of s_{33} over a prescribed band of frequencies subject to the general constraints of physical realizability as well as the requirements on the behavior of s_{33} at $p = \infty$. First we shall assume that $\omega_s\omega_u \cong \omega_{so}\omega_{uo}$ over the frequency band of interest where ω_{so} and ω_{uo} are the respective band center frequencies, and that $Y_3(j\omega_s) \cong Y_3(j\omega_u)$, which can be realized approximately by using a low-pass to multiple band-pass frequency transformation. With these two assumptions, Equations (1) and (7) become respectively

$$z_D(j\omega_s) = \frac{1}{\omega_{so}\omega_{uo}C_1^2 Z_3(j\omega_s)} \quad , \qquad (9)$$

and

$$s_{33}(j\omega_s) = \frac{\omega_{so}\omega_{uo}C_1^2 - |Y_3(j\omega_s)|^2}{\omega_{so}\omega_{uo}C_1 + Y_3^2(j\omega_s)} = \frac{1 - |Y(j\omega_s)|^2}{1 + Y^2(j\omega_s)} \quad , \tag{10}$$

where $Y \equiv Y_3 / \sqrt{\omega_{so}\omega_{uo}}\, C_1$. Now, the real part of z_D is non-negative on the entire $j\omega_s$ axis. It follows that $|s_{33}(j\omega_s)|^2 \leq 1$, and

$$G_t(\omega_s) = \frac{\omega_u}{\omega_s}(1 - |s_{33}(j\omega_s)|^2) \leq \frac{\omega_u}{\omega_s} \tag{11}$$

for all ω_s. Substitute Equation (10) in (11) and let $G(\omega_s^2)$ be defined as $G(\omega_s^2) \equiv (\omega_s/\omega_u)G_t(\omega_s)$, then

$$\left(\frac{\omega_s}{\omega_u}\right) G_t(\omega_s) = G(\omega_s^2) = 1 - |s_{33}(j\omega_s)|^2 = 1 - \left\{\frac{\left[1 - |Y(j\omega_s)|^2\right]^2}{|1 + Y^2(j\omega_s)|^2}\right\} \leq 1. \tag{12}$$

Hence, $G(\omega_s^2)$ is an even function of ω_s which is bounded by unity on the entire $j\omega_s$ - axis.

To proceed with the broadbanding problem, we consider the analytic extensions of the steady-state expressions to the entire p-plane. From Equations (10) and (12),

$$s_{33}(p) = \frac{1 - Y(p)Y(-p)}{1 + Y^2(p)} \tag{13}$$

and

$$G(-p^2) = 1 - s_{33}(p)s_{33}(-p) = \frac{[Y(p) + Y(-p)]^2}{[1 + Y^2(p)]\,[1 + Y^2(-p)]} \quad . \tag{14}$$

Under the matched condition, $s_{33}(p) = 0$, and $G(-p^2) = 1$. Thus

$$Y(p)Y(-p) \equiv \frac{Y_3(p)Y_3(-p)}{\left(\sqrt{\omega_{so}\omega_{uo}}\, C_1\right)^2} = 1 \quad , \tag{15}$$

and the problem is immediately reduced to that of equalization of a 1-port shunted by a capacitance. This condition, however, is not imposed in this paper.

From Equation (14), we see that a realizable transducer power gain $G_t(p) = (1 + j\omega_o/p)G(-p^2)$ must be of such a form that

$G(-p^2)/1 - G(-p^2)$ is a perfect square of a real, rational, even function of p since

$$\frac{G(-p^2)}{1 - G(-p^2)} = \left[\frac{Y(p) + Y(-p)}{1 - Y(p)Y(-p)}\right]^2 \quad . \tag{16}$$

Therefore,

$$\sqrt{\frac{G(-p^2)}{1 - G(-p^2)}} = \pm\left[\frac{Y(p) + Y(-p)}{1 - Y(p)Y(-p)}\right] \quad . \tag{17}$$

Choosing the minus sign, Equation (17) can be expressed as

$$\sqrt{\frac{G(-p^2)}{1 - G(-p^2)}} = -\left[\frac{1 + Y(p)Y(-p)}{1 - Y(p)Y(-p)}\right]\left[\frac{Y(p) + Y(-p)}{1 + Y(p)Y(-p)}\right] \; . \tag{18}$$

In order to have a reasonable amount of gain, $s_{33}(p)$ cannot be too close to unity in the frequency band of interest. Hence, from Equation (18), the variation of the second factor is small as compared to that of the first and can be neglected. By imposing this approximation, Equation (18) becomes

$$\sqrt{\frac{G(-p^2)}{1 - G(-p^2)}} \cong -\left[\frac{1 + Y(p)Y(-p)}{1 - Y(p)Y(-p)}\right] \; , \tag{19}$$

and

$$\frac{\sqrt{\frac{G(-p^2)}{1 - G(-p^2)}} + 1}{\sqrt{\frac{G(-p^2)}{1 - G(-p^2)}} - 1} = Y(p)Y(-p) \quad , \tag{20}$$

where $Y(p) \equiv Y_3(p)\Big/\sqrt{\omega_{so}\omega_{uo}}\, C_1$ is positive-real with a prescribed residue at $p = \infty$ due to the parasitic capacitance of the varactor. More precisely, at $p = \infty$,

$$Y(p) = pC + y(p) \quad , \tag{21}$$

where C is the total effective shunt capacitance, $C \geq C_o\Big/\sqrt{\omega_{so}\omega_{uo}}\, C_1$, and y(p) is an arbitrary positive-real function which is finite at $p = \infty$.

For $p = j\omega$ and $Z(j\omega) = 1/Y(j\omega)$, Equation (20) gives

$$\frac{\sqrt{\frac{G(\omega^2)}{1 - G(\omega^2)}} - 1}{\sqrt{\frac{G(\omega^2)}{1 - G(\omega^2)}} + 1} = |Z(j\omega)|^2 \quad . \tag{22}$$

Applying Bode's well-known integral restriction(5) to $\ln Z(j\omega) = \ln |Z(j\omega)| + j\theta(j\omega)$, we have

$$\int_0^{\omega_c} \frac{\ln|Z(j\omega)|}{\sqrt{1 - \frac{\omega^2}{\omega_c^2}}} d\omega = \int_{\omega_c}^{\infty} \frac{\theta(j\omega)}{\sqrt{\frac{\omega^2}{\omega_c^2} - 1}} d\omega \quad . \tag{23}$$

Bode further established that the maximum constant gain for a 1-port interstage shunt by a capacitance C is given by(5)

$$|Z(j\omega)|_{max} = \frac{2}{\omega_c C} , \tag{24}$$

where ω_c is the low-pass bandwidth, defined as the frequency band through which the gain of the interstage is constant. Let the bandwidths of the signal and the upper-sideband frequency pass bands be denoted by B. From the conservation of bandwidth as applied to the low-pass to double bandpass transformation, it is clear that B is equal to one-half of the available low-pass bandwidth, ω_c.

Applying Bode's results to Equation (22) and noting that $B = \omega_c/2$, we obtain

$$\left[\frac{\sqrt{\frac{G_{max}}{1 - G_{max}}} - 1}{\sqrt{\frac{G_{max}}{1 - G_{max}}} + 1}\right]^{1/2} = \frac{2}{\omega_c C} = \frac{1}{BC} \leq \frac{\sqrt{\omega_{so}\omega_{uo}}\, C_1}{BC_o} \quad , \tag{25}$$

where $G_{t_{max}} = (\omega_u/\omega_s)\, G_{max} \leq (\omega_u/\omega_s)$ and the equality is attained if the equalizer does not contribute additional effective capacity, i.e., $C = C_o/\sqrt{\omega_{so}\omega_{uo}}\, C_1$. The relation (25) contains the fundamental gain-bandwidth limitation of the parametric up-converter which is applicable in general irrespective of whether the up-converter circuit is matched or not.* It can be concluded from relation (25) that for a parametric up-converter mismatch can actually be traded with bandwidth, i.e., by decreasing G_{max}, (thus increasing mismatch), the bandwidth is increased. This fundamental limitation is plotted in Figure 4, with the curve labelled $n = \infty$. The dotted portion signifies the fact that the limitation is approximate when the mismatch is large (see Equations (18) and (19)).

For the special case of the matched condition, $G_{max} = 1$, and Equation (25) reduces to

$$B = \frac{\omega_c}{2} = \frac{\sqrt{\omega_{so}\omega_{uo}}\, C_1}{C_o} \quad , \tag{26}$$

which is the limitation derived previously by Kuh.(4)

4. MAXIMALLY FLAT PARAMETRIC UP-CONVERTERS

We have presented in the preceding section a gain-bandwidth limitation for the parametric up-converter which is independent of the equalizer. It is to be noted, however, that the resulting optimum equalizer impedance for a prescribed G_{max} is not rational(5) and some approximation is necessary. In this section, we shall derive some explicit results for lossless low-pass ladder equalizers with finite number of elements. These equalizers have the additional property that $G(\omega_s^2) = (\omega_s/\omega_u)G_t(\omega_s)$ is maximally flat.

Let $Y(p) = N(p)/D(p)$, where N and D are polynomials, then from Equation (14) we have

$$G(-p^2) = \frac{\left[N(p)D(-p) + N(-p)D(p)\right]^2}{\left[D(p)D(-p) - N(p)N(-p)\right]^2 + \left[N(p)D(-p) + N(-p)D(p)\right]^2} \quad , \tag{27}$$

where the polynomial $\left[N(p)D(-p) + N(-p)D(p)\right]$ is precisely the numerator of the even part of $Y(p)$. Consider a lossless low-pass ladder of order n terminated in a conductance G_n (Figure 3), then the transmission zeros must all be at $p = \infty$. It follows that $\left[N(p)D(-p) + N(-p)D(p)\right]$ is devoid of zeros in the entire p-plane and is therefore a constant. From Figure 3, it is clear that this constant is equal to $2G_n$. Then Equation (27) becomes

---

*Note that (25) is not valid for mismatch loss exceeding 3 db since it predicts that $\omega_c \rightarrow \infty$ as $G_{max} \rightarrow 1/2$, which is impossible. This results from the approximation embodied in (19).

$$G(-p^2) = \frac{4G_n^2}{\left[D(p)D(-p) - N(p)N(-p)\right]^2 + 4G_n^2} \ . \tag{28}$$

Represent the even polynomial $\left[D(p)D(-p) - N(p)N(-p)\right]$ for $p = j\omega$ by

$$\omega^{2n}(\alpha_n) + \omega^{2(n-1)}(\alpha_{n-1}) + \cdots + \omega^2(\alpha_1) + \alpha_o \ ,$$

where $\alpha_o = (G_n^2 - 1)$ and $\alpha_n = (CL \cdots L_{n-1} \, C_n)^2$ for n odd and $\alpha_n = (CL_2 \cdots C_{n-1} L_n G_n)^2$ for n even. From Figure 3, we see that C is prescribed by the fixed parasitic capacitance and G_n is prescribed by the d.c. gain of the up-converter. There remains (n-1) parameters representing the (n-1) reactances at our disposal. For maximally flat response,(9,10) we can impose the conditions that

$$\alpha_r = 0 \ , \quad r = 1,2,\cdots,n-1 \ , \tag{29}$$

which comprises (n-1) simultaneous equations. For maximally flat response shape, we have

$$G(\omega^2) = \frac{4G_n^2}{(\alpha_n \omega^{2n} + \alpha_o)^2 + 4G_n^2} \tag{30}$$

or, in terms of $\gamma_n \equiv \sqrt{\alpha_n}$,

$$G(\omega^2) = \frac{\dfrac{4G_n^2}{\gamma_n^4}}{\omega^{4n} + \left[2\,\dfrac{(G_n^2 - 1)}{\gamma_n^2}\right]\omega^{2n} + \dfrac{(G_n^2 + 1)^2}{\gamma_n^4}} \ . \tag{31}$$

The d.c. gain G(o) and G_n are related by

$$G(o) = \frac{4G_n^2}{(G_n^2 + 1)^2} \leq 1 \quad , \tag{32}$$

and

$$G_n = \frac{1}{\sqrt{G(o)}}\left[1 + \sqrt{1 - G(o)}\right] \geq 1 \quad . \tag{33}$$

The 3 db bandwidth of the signal and the upper-sideband passbands, B, can be derived from Equation (31) as

$$B = \frac{\omega_c}{2} = \frac{1}{2\gamma_n^{1/n}}\left\{\sqrt{2\left[\left(\frac{1 + \sqrt{1 - G(o)}}{\sqrt{G(o)}}\right)^4 + 1\right]} - \left[\left(\frac{1 + \sqrt{1 - G(o)}}{\sqrt{G(o)}}\right)^2 - 1\right]\right\}^{\frac{1}{2n}} \tag{34}$$

where G(o) is the d.c. gain and γ_n can be expressed in terms of G(o) and $C = C_o / \sqrt{\omega_{so}\omega_{uo}}\, C_1$ by utilizing conditions (29) for maximal flatness. Note that in the above derivation, we did not make use of the approximation emboded in (19), which is used to derive the optimum gain-bandwidth limitation. For the special case under the matched condition, G(o) = 1 and the relation (34) simplifies to

$$B = \frac{\omega_c}{2} = \frac{1}{2}\left(\frac{\sqrt{2}}{\gamma_n}\right)^{1/n} \quad . \tag{35}$$

Explicit analytic results for bandwidth and element values are presented in the Appendix for equalizers of order n = 1, 2, and 3, which correspond respectively to single-tuned, double-tuned, and triple-tuned maximally flat up-converters. For n $\geq$ 4, explicit results can be obtained by solving simultaneous nonlinear algebraic equations on a computer. These results are plotted in Figure 4 to facilitate the design of broadband parametric up-converters. It is not possible to generalize the explicit gain-bandwidth limitations of the parametric device for an arbitrary order n using the procedure outlined by Sard,$^{(10)}$ who treated tunnel-diode amplifiers with maximally flat responses.

$$B = \frac{\omega_c}{2} = \frac{\sqrt{\omega_{so}\omega_{uo}}\, C_1}{2C_o} \left\{ \frac{2(2 + \sqrt{G_3^2 + 1}\,\beta)}{\left(\dfrac{G_3}{1 + \sqrt{G_3^2 + 1}\,\beta}\right)^2 - 1} \right\}^{\frac{1}{3}} \cdot$$

$$\left\{ \sqrt{2\left[\left(\frac{1 + \sqrt{1 - G(o)}}{\sqrt{G(o)}}\right)^4 + 1\right]} - \left[\left(\frac{1 + \sqrt{1 - G(o)}}{\sqrt{G(o)}}\right)^2 - 1\right] \right\}^{\frac{1}{6}} \quad ; \qquad (A.6)$$

where

$$\beta = \sqrt{3}\,\sin\frac{\phi}{3} - \cos\frac{\phi}{3}\,, \qquad \phi = \cos^{-1}\frac{1}{\sqrt{G_3^2 + 1}}\,, \qquad (A.7)$$

and G_3 and $G(o)$ are related by Equation (33).

$n \geq 4$

For the case of n equal to or greater than four, it is still possible to obtain explicit solutions for the element values and the gain-bandwidth product. But, now, it is necessary to solve for the element values from the (n - 1) maximally flatness conditions on a computer for a specified gain G(o). For example, for n = 4, the maximally flat conditions to be imposed are $\alpha_3 = 0$, $\alpha_2 = 0$ and $\alpha_1 = 0$ which result in a set of three simultaneous equations from which we can solve for L_2, C_3, and L_4. For $G(o) = 0$, we have

$$L_2 = 0.848\ C, \quad C_3 = 0.566\ C, \quad L_4 = 0.199\ C\ , \qquad (A.8)$$

and

$$B = \frac{\omega_c}{2} = (0.983)\,\frac{\sqrt{\omega_{so}\omega_{uo}}\, C_1}{C_o}\ . \qquad (A.9)$$

Matched Up-Converters

It is interesting to note that under the matched condition, $G(o) = 1$, the maximally flat parametric up-converters using lossless ladders represent a very good approximation to the optimum bandwidth of a matched up-converter, $B = \sqrt{\omega_{so}\omega_{uo}}\, C_1/C_o$, first derived by Kuh.

TABLE 1

BANDWIDTH LIMITATIONS AND ELEMENT VALUES FOR A MATCHED EQUALIZER OF ORDER n

	$B \cdot (C_o / \sqrt{\omega_{so}\omega_{uo}}\; C_1$	L_2/C	C_3/C	L_4/C
$n = 1$	0.707	-	-	-
$n = 2$	0.925	0.414	-	-
$n = 3$	0.966	0.732	0.268	-
$n = 4$	0.983	0.848	0.566	0.199
$n = \infty$	1.000	-	-	-

The results for the matched case are tabulated in Table 1, where B is the 3 db bandwidth and $C = C_o / \sqrt{\omega_{so}\omega_{uo}}\; C_1$. Note also that $G(o) = 1$ and $G_t(o) = \omega_u/\omega_s$ for the matched case.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Dr. P. R. Johannessen, of the Applied Research Laboratory, Sylvania Electronic Systems, and Dr. A. L. Helgesson of the Melabs, Palo Alto, California, for stimulating discussions in the course of preparing this paper.

REFERENCES

1. J. M. Manley and H. E. Rowe, "Some General Properties of Nonlinear Elements, Part I. General Energy Relations," Proc. IRE, Vol. 44, No. 7 (July 1956), pp. 904-913.

2. H. Seidel and G. Herrman, "Circuit Aspects of Parametric Amplifiers," IRE WESCON Convention Record, Part 2 (1959), pp. 83-90.

3. G. L. Matthaei, "A Study of the Optimum Design of Wide Band Parametric Amplifiers and Up-Converters," IRE Trans. on Microwave Theory and Techniques, Vol. MTT-9, No. 1 (January 1961), pp. 23-38.

4. E. S. Kuh, "Theory and Design of Wide-Band Parametric Converters," Proc. IRE, Vol. 50, No. 1 (January 1962), pp. 31-38.

5. H. W. Bode, Network Analysis and Feedback Amplifier Design, D. VanNostrand, Inc., 1945.

6. A. L. Helgesson, "Gain-Bandwidth Limitations of Balanced Three-Frequency Parametric Amplifiers," Ph.D. Dissertation, MIT, 1962.

7. D. C. Youla, "On Scattering Matrices Normalized to Complex Port Numbers," Proc. IRE, Vol. 49, No. 7 (July 1961), p. 1221.

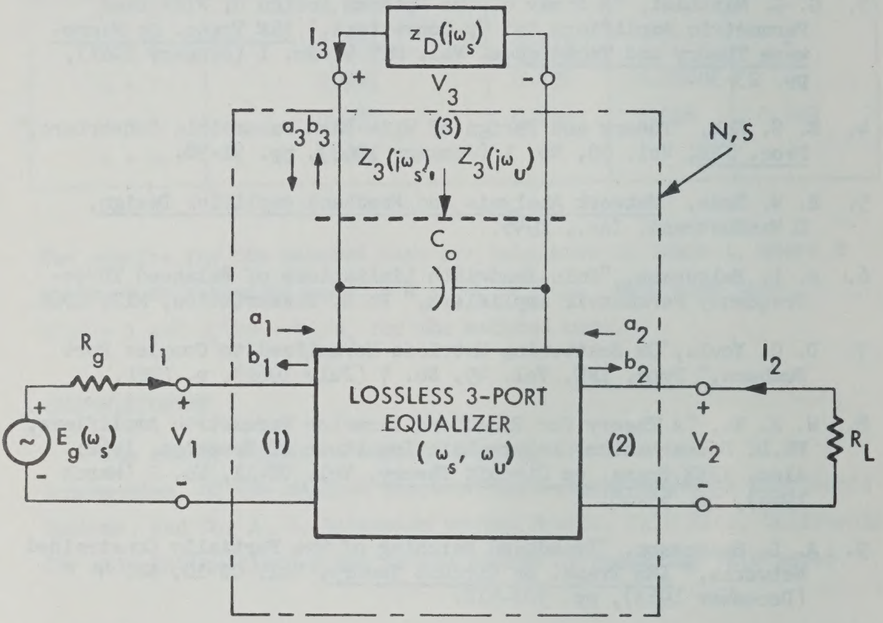
8. W. H. Ku, "A Theory for Broadband Varactor Parametric Amplifiers," Ph.D. Dissertation, Polytechnic Institute of Brooklyn, 1962; Also, IEEE Trans. on Circuit Theory, Vol. CT-11, No. 2 (March 1964), pp. 53-68.

9. A. L. Helgesson, "Broadband Matching of Two Partially Constrained Networks," IRE Trans. on Circuit Theory, Vol. CT-10, No. 4 (December 1963), pp. 503-512.

10. W. H. Ku, "Broadband Equalization of Two Partially Constrained Passive Networks," IRE Trans. on Circuit Theory, Vol. CT-10, No. 4 (December 1963), pp. 538-539.

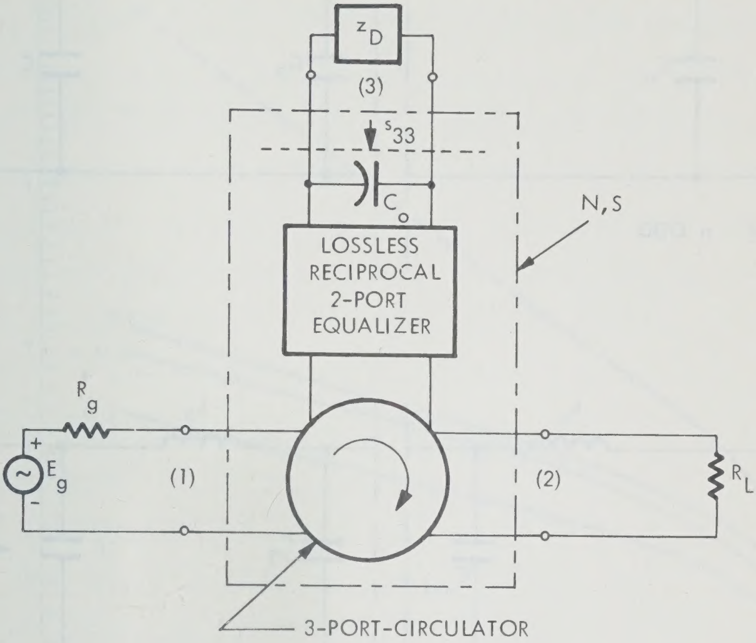
11. W. A. Lynch, "The Role Played by Derivative Adjustment in Broadband Amplifier Design," Proc. Symp. on Modern Network Synthesis, Polytechnic Institute of Brooklyn, April 1952.

12. E. W. Sard, "Gain-Bandwidth Performance of Maximally Flat Negative-Conductance Amplifiers," Proc. Symp. on Active Networks and Feedback Systems, Polytechnic Institute of Brooklyn, April 1960.


$z_D(j\omega_s)$
I_3
+
V_3
-
$a_3 b_3$
(3)
N, S
$Z_3(j\omega_s)$, $Z_3(j\omega_u)$
C_o
a_1
b_1
a_2
b_2
R_g
I_1
I_2
LOSSLESS 3-PORT EQUALIZER
(ω_s, ω_u)
+
-
$E_g(\omega_s)$
V_1
(1)
(2)
V_2
R_L

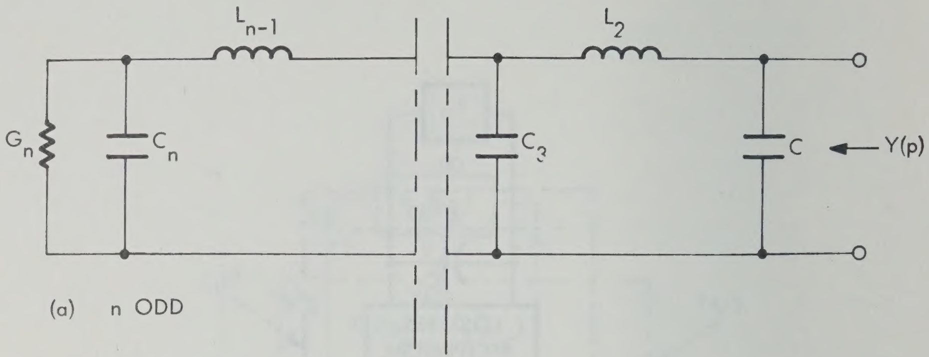
$$z_D(j\omega_s) = \frac{1}{\omega_s \omega_u C_1^2 \, Z_3(j\omega_u)} ; \quad \underset{\sim}{b} = S\underset{\sim}{a}$$

Figure 1. General Configuration of Varactor Parametric Up-Converter

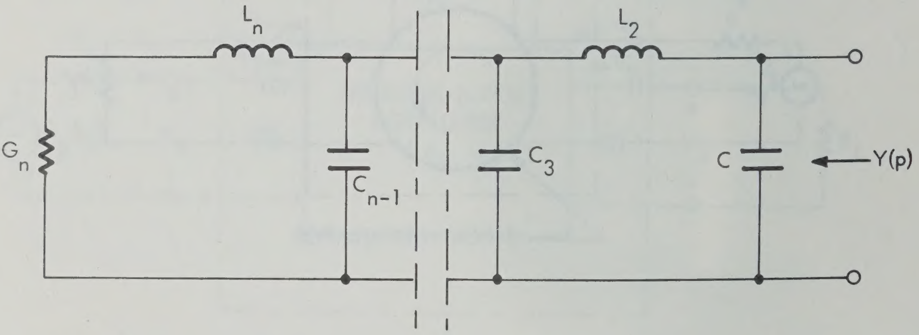


z_D
(3)
s_{33}
C_o
N, S
LOSSLESS RECIPROCAL 2-PORT EQUALIZER
R_g
+
E_g
-
(1)
(2)
R_L
3-PORT-CIRCULATOR

Figure 2. Varactor Parametric Up-Converter Using a Three-Port Circulator


L_{n-1}
L_2
G_n
C_n
C_3
C
Y(p)

(a) n ODD


L_n
L_2
G_n
C_{n-1}
C_3
C
Y(p)

(b) n EVEN

Figure 3. Lossless Low-Pass Ladders of Order n. The Capacitance $C = C_o / \sqrt{\omega_{so}\omega_{uo}}\ C_1$ is Prescribed by the Varactor

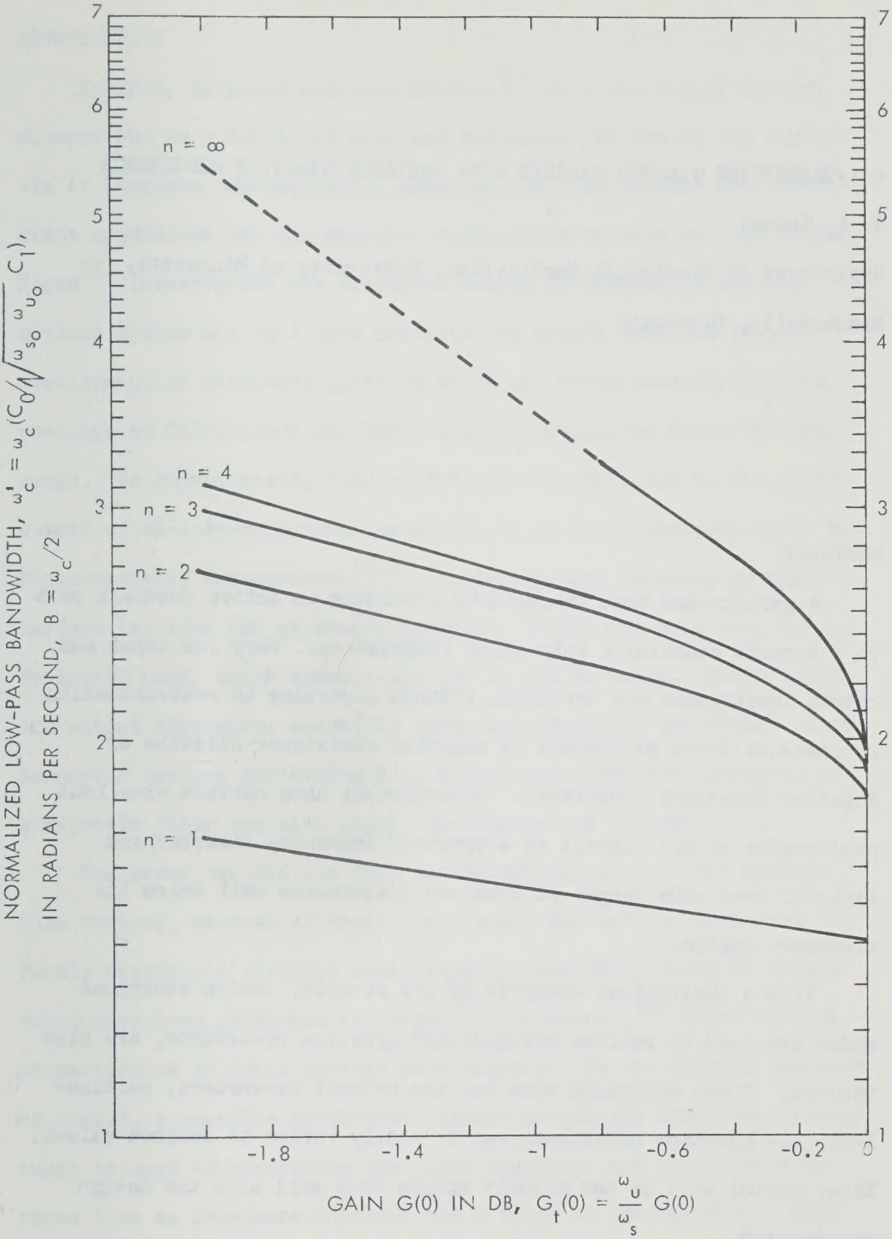


Figure 4. Gain-Bandwidth Relations for Non-Inverting Parametric Up-Converters

A 3-TRANSISTOR GYRATOR CIRCUIT WITH VARIABLE GYRATION RESISTANCE

B. A. Shenoi

Department of Electrical Engineering, University of Minnesota,
Minneapolis, Minnesota

ABSTRACT

A gyrator has been designed by providing an active feedback path in a circuit containing only three transistors. Very low input and output immittances are obtained, without resorting to neutralization of residual input parameters by negative resistance circuits or Negative Impedance Converters. Experimental data exhibit excellent performance of the circuit as a Gyrator, Impedance Inverter, and Isolator over wide ranges of load and frequencies well below the microwave region.

From a theoretical analysis of the circuit, design equations which are used to realize prespecified gyration resistance, are also derived. These equations show how the network parameters, particularly the gyration resistance, can be easily varied to desired values. Experimental work on the circuit agrees very well with the design theory also.

INTRODUCTION

In 1948, Tellegan proposed Gyrator$^{(1)}$ as a new 2-part network element and in a series of articles discussed its use in the synthesis of lossless non-reciprocal networks and also derived some sufficient conditions for the physical realization of this device. Later Hogan$^{(2)}$ investigated the microwave analog of Rayliegh's one-way optical system and this work initiated an effort that led to the development of microwave gyrators which are being extensively used nowadays as Circulators and Isolators at frequencies above the KMc range. In recent years, some effort has been directed to the development of electro-mechanical gyrators, by coupling piezo-electric and piezo-magnetic transducers.$^{(3,4)}$ Antireciprocal behavior of these devices is, like the microwave gyrators, restricted to a very narrow frequency band, which however may lie in the Kc range. Their input and output impedances are fairly large and frequency-dependent. Semiconductor devices exhibiting Hall Effect when they are subjected to a magnetic field are also being investigated at present.

The above devices are only antireciprocal and are not gyrator-like devices, because of their large input and output impedances. Purely electronic circuits exploiting the non-reciprocity of vacuum tubes have been developed by Sharpe$^{(5)}$ and Bogert,$^{(6)}$ which exhibit properties of an ideal gyrator more closely. In the circuit designed by Bogert, a negative resistance circuit containing additional vacuum tubes is used to neutralize the input impedance and the circuit behaves like an Impedance Invertor for a range of load $40 \leqslant R_L \leqslant 4000$ ohms. Last year, Ghausi and McCarthy reported a gyrator circuit containing transistors.$^{(7)}$ They used two negative resistance circuits to neutralize the driving point impedances at both the input and output terminals, which were of the same magnitude as the transfer impedances of the basic antireciprocal circuit. They used two Negative Impedance Converters (NIC) for obtaining these negative resistances and the total circuit had six transistors and a large number of passive elements, though the realization of the of the 2-port was conceptually simple.*

CIRCUIT ANALYSIS AND DESIGN THEORY OF THE GYRATOR

A gyrator circuit using only three transistors has been realized by this author and its ac equivalent circuit is shown in Figure 1. It is presented in the form of 2 networks N_a and N_b connected in parallel. The s.c. admittance matrix of an ideal gyrator is

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} 0 & -G_o \\ +G_o & 0 \end{bmatrix} \tag{1}$$

from which the o.c. impedance matrix can also be derived as

$$\begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{y_{21}} \\ \frac{1}{y_{12}} & 0 \end{bmatrix} = \begin{bmatrix} 0 & R_o \\ -R_o & 0 \end{bmatrix} \tag{2}$$

where $R_0 = 1/G_0$ is real and known as the Gyration Resistance. The input and output immittances of the ideal gyrator are zero, its 'feedback' is as large as the 'forward gain' in magnitude but opposite in phase. By drawing a signal flow graph or an equivalent single loop feedback diagram, it can be shown that the return difference is infinite and $z_{12} = 1/y_{21}$, $z_{21} = 1/y_{12}$. All these conditions are approached very closely by the gyrator circuit shown in Figure 1, which has an active feedback path to provide a gain (= current gain of N_b x Load resistance of N_b) that is adjusted to be equal to the forward gain (= current gain of N_a x Load resistance of N_a) in magnitude but opposite in phase. Its return difference is very large, input and output impedances are found to be very low due to the shunt feedback and the large return difference. Hence, unlike the circuits developed by the earlier authors,(6,7) no additional negative resistance circuits are deemed necessary for this circuit.

---

*The present author has designed a circuit in which only one NIC is used to neutralize the residual output impedance, the input impedance of the network being very low. This circuit has only four transistors and can realize a wide range of Gyration Resistance.

The s.c. admittances of the gyrator are derived in Appendix A and the o.c. impedances obtained under the assumption $y_{11}y_{22} - y_{12}y_{21} = - y_{12}y_{21}$ are given by Equations (A.15), (A.17), (A.19) and (A.22). For purposes of preliminary design, further assumptions are made that in (A.17) $\alpha_3 \approx 1$ and $z_{b3} \ll r_{b3}$ so that R_{b3} can be neglected. In Equation (A.19) the term $g_{b2}C$ is also neglected. The above equations are thus reduced to Equations (3) through (6) in which $r_{ej} + \frac{r_{bj}}{\beta_j} = \frac{h_{iej}}{\beta_j}$.

$$z_{12} \approx R_{e1} + \frac{h_{ie1}}{\beta_1} \tag{3}$$

$$z_{21} \approx - \frac{h_{ie2}}{\beta_2}\left(\frac{h_{ie3}}{\beta_3 R_5} + 1\right) - \frac{h_{ie3}}{\beta_3}\left(\frac{R_4}{R_5} + 1\right) - R_4$$

$$\approx -R_4\left(1 + \frac{h_{ie3}}{\beta_3 R_5}\right) - \frac{h_{ie3}}{\beta_3}\left(1 + \frac{h_{ie2}}{\beta_2 R_5}\right) - \frac{h_{ie2}}{\beta_2} \tag{4}$$

$$z_{11} \approx -z_{12}z_{21}(G_{L1}+G_{b2}) = R_o^2(G_{L1}+G_{b2}) \qquad (5)$$

$$z_{22} \approx -z_{12}z_{21}(G_{b1}+G_{L3}) - \frac{z_{21}}{\beta_1} \qquad (6)$$

Now h_{iej} is the input impedance of a CE transistor T_j when its collector is shorted and is quantity that can be obtained from the specification sheet or by a quick measurement of the individual transistors chosen for use. Many high gain transistors are available for which $\frac{h_{iej}}{\beta_j}$ has a constant value of about 10 or less when $I_E > 5$ mA. Hence by making R_{e1} and R_5 large, it can be seen from Equations (3) and (4) that the effect of this transistor parameter on z_{12} and z_{21} can be made small.

To illustrate the application of the above design equations, let the specified z-matrix parameters of the gyrator by $z_{11} = 0.7$, $z_{12} = 120$, $z_{21} = -120$ and $z_{22} = 2.0$. Beta cut-off frequency of transistors T_1 and T_3 is 4 Mc and that of T_2 is about 500 Mc. From actual measurements, it is found that

$$\beta_1 = 100 \; ; \; \frac{h_{ie1}}{\beta_1} = 10 \text{ at } I_E = 5.4 \text{ mA and } V_{CE} = 14 \text{ V} \qquad (7)$$

$$\beta_2 = 2000 \; ; \; \frac{h_{ie2}}{\beta_2} = 10 \text{ at } I_E = 5 \text{ mA and } V_{CE} = 11 \text{ V} \qquad (8)$$

$$\beta_3 = 100 \; ; \; \frac{h_{ie3}}{\beta_3} = 10.5 \text{ at } I_E = 5.2 \text{ mA and } V_{CE} = 8 \text{ V} \qquad (9)$$

From Equations (3) and (7), $R_{e1} = 110\ \Omega$. R_5 is chosen $130\ \Omega$. From Equations (4), (8) and (9) R_4 is therefore $91\ \Omega$. To get $z_{11} = 0.7$, Equation (5) requires that $(G_{L_1} + G_{b_2}) = .486 \times 10^{-4}$. G_{L_1} is chosen to be $.37 \times 10^{-4}$, then $G_{b_2} = .116 \times 10^{-4}$ which is the total conductance of the two biasing resistors R_{v3} and R_{v4} shown in Figure 4. (However values of these resistors have been actually increased in order to allow for the term $g_{b_2}C$ which is included in the value $.116 \times 10^{-4}$.)

From Equation (6), we find $(G_{b_1} + G_{L_3}) = .555 \times 10^{-4}$ is needed to get $z_{22} = 2.0$. Choosing $G_{L_3} = 0.37 \times 10^{-4}$ ($R_{L_3} = 27$ KΩ) requires that $G_{b_1} = .185 \times 10^{-4}$ which is the total conductance of the two parallel resistances R_{v1} and R_{v2} used for biasing T_1. A circuit constructed according to the above design gives the following results: $z_{11} = .685$, $z_{12} = 116$, $z_{21} = -118$ and $z_{22} = 2.2$. This agrees very closely with the specifications assumed. By adding a small resistance to R_4, z_{12} is adjusted to $+118$ to get a z-matrix

$$\begin{bmatrix} 0.685 & 118 \\ -118 & 2.2 \end{bmatrix}$$

The final circuit is shown in Figure 4. (z_{12} and z_{21} of this circuit could have been easily adjusted to design values, but further experimental data was obtained without making this adjustment.) The rest of the biasing circuitry is also included in it. Usually in the analysis and synthesis or design of transistor circuits the biasing elements are completely omitted, on the assumption that their effect on ac performance can be neglected — which is not always true. In the present case the effect of the biasing resistances R_{v_1}, R_{v_2}, R_{v_3} and R_{v_4} which are effectively in parallel with the load resistances R_{L_1} and R_{L_3} is considered in the design of the circuit, as is illustrated in the above calculations. The conductances are so distributed among the load and the bias resistances that the latter establish proper dc bias conditions also.

EXPERIMENTAL RESULTS

The above values for the z parameters are those of the circuit shown in Figure 4, at 1 Kc. The variation of the z parameters with frequency is shown in Figures 5 and 6. They are constant over nearly 2.5 decades of frequency except z_{22}. The deviation of z_{22} is mainly due to the shunt capacitance C_1 in Figure 4; otherwise z_{22} would have a characteristic similar to z_{11}. The driving point impedance characteristics of the gyrator when it is loaded by different resistances at the output is shown in Figure 7 and it shows that the gyrator behaves like an ideal Impedance Inverter over a 2.5 decade range of load (e.g., from 25 to 7000 ohms). Figure 8 shows the input impedance characteristics as the frequency is varied when the load is a fixed capacitance of .133 μf.

When a conductance equal in magnitude to y_{12} of the gyrator is connected in parallel with the circuit, the reverse transfer admittance of the resulting circuit is neutralized and this unilateral circuit can be used as an Isolator. Measurements show that the input driving point impedance is now the same for all load resistances and the reverse transfer impedance remains constant over a wider frequency range than that of Figure 5. The characteristics of the Isolator are obtained by measuring the insertion loss in the forward (isolating) and reverse (transmitting) direction. From Figure 9 it is seen that the unilateralized circuit has a maximum insertion loss of 60 db in the isolating direction as compared to 1.15 db in the transmitting direction. The insertion loss in the isolating direction is constant within 10% over a wide bandwidth of 2 Kc and 250 Kc, whereas that in the transmitting direction has a still wider bandwidth. Using better transistors and biasing arrangement, the gyrator can be further improved but from a comparison of its characteristics with the ideal characteristics which are plotted in Figures 5, 6, 7, and 8 it is evident that over a 2.5 decade of frequency and load, the experimental results coincide very well with the theoretical predictions. The deviations at the higher and lower frequencies are due to the coupling capacitors in the circuit and those at the extremes of the load range are of course due to the non-zero values of the residual driving point impedances. From the expression

$$z_{in} = z_{11} - \frac{z_{12}z_{21}}{z_{22} + z_L}$$

it is clear that as the product $(-z_{12}z_{21}) = R_0^2$ is increased, the effect of z_{11} on the deviation from the ideal impedance inversion characteristic is decreased. But in this gyrator circuit, as seen from Equations (5) and (6), z_{11} and z_{22} both increase slowly as $(-z_{12}z_{21})$ is increased. Hence a mere increase in $(-z_{12}z_{21})$ does not considerably extend or change the range of load over which the circuit behaves like an ideal Impedance Invertor but only shifts the range as seen from curves in Figure 7. But the design Equations (3) and (4) confirm our experience that z_{12} and z_{21} can be individually varied by simply adjusting R_{e_1} and R_4. The simplicity and extent of this control is a significant feature of the device, so that the same circuit can be employed when different Gyration Resistance R_0 or Inversion Factor R_0^2 are desired. The several curves in Figure 7 are obtained merely by varying R_{e_1} and R_4. The alterations needed in the circuits developed by the earlier authors, to change the Gyration Resistance call for a new design every time.

The input and output impedances, however, can also be reduced by increasing the load resistances and also the biasing resistances connected in parallel. In order to establish proper biasing currents and voltages, it therefore becomes necessary to use higher bias supply voltages for this purpose. The resistances R_8, R_9 (and C_9) are used to provide a dc feedback that helps to stabilize the bias conditions for transistors T_1 and T_2 and thus suppress any tendency for slow drifting of the output response. The tendency for unstable oscillations in any gyrator circuit is greatest when the output is open circuited since then the input impedance seen by the generator is the lowest. But the circuit in Figure 4 has been found to be open circuit stable for all frequencies up to 500 Kc, when C_1 = .05 μf; however, the value of C_1 = 400 μμf is found sufficient to assure stability up to 500 Kc when it is loaded or shorted at the output.

APPENDIX A

The T-equivalent circuit assumed for the common base transistor T_j is shown in Figure 1.b; those for CE and CC configuration being obtained in terms of the parameters of this circuit. The reasons for this choice are that (1) the z parameters of the gyrator must be real, if a passive gyrator is sought to be realized (2) the range of frequencies over which the circuit is used as a gyrator is much below any inverse time constant of the network so that the effect of all capacitances can be neglected (3) the complexity of network analysis is reduced and the last but not the least important is the fact that (4) the experimental results amply justify the validity of the theoretical predictions obtained by assuming such a simple equivalent circuit.

From the mode admittance matrix $[Y]$ relating $[I_1 \;\; I_2 \;\; 0]^T$ and $[E_1 \;\; E_2 \;\; E_3]^T$, the s.c. admittances of network N_a shown in Figure 2 are derived as follows:

$$Y_{11a} = \frac{\Delta_{22}}{\Delta_{1122}} = G_{b1} + \frac{g_{b1}(G_{E1} + g'_{e1})}{g_{b1} + \beta_1 g_{b1} + G_{E1} + g'_{e1}} \tag{A.1}$$

$$Y_{12a} = \frac{-\Delta_{21}}{\Delta_{1122}} = \frac{-g_{b1} g'_{e1}}{g_{b1} + \beta_1 g_{b1} + G_{E1} + g'_{e1}} \tag{A.2}$$

$$Y_{21a} = \frac{-\Delta_{12}}{\Delta_{1122}} = \frac{g_{b1}(\beta_1 G_{E1} - g'_{e1})}{g_{b1} + \beta_1 g_{b1} + G_{E1} + g'_{e1}} \tag{A.3}$$

$$Y_{22a} = \frac{\Delta_{11}}{\Delta_{1122}} = G_{L1} + \frac{g'_{e1}(G_{E1} + g_{b1})}{g_{b1} + \beta_1 g_{b1} + G_{E1} + g'_{e1}} \tag{A.4}$$

Similarly those for network N_b are obtained from Figure 3:

$$Y_{11b} \simeq G_{L3} \tag{A.5}$$

$$Y_{12b} = \frac{-\alpha_3}{\left(r_{e2} + \frac{r_{b2}}{\beta_2}\right)\left[\left(\frac{R_{b3}}{\beta_3} + \frac{r_{b3}}{\beta_3} + r_{e3}\right)\frac{1}{R_5} + 1\right] + \left[\left(1 + \frac{R_4}{R_5}\right)\left(\frac{R_{b3}}{\beta_3} + \frac{r_{b3}}{\beta_3} + r_{e3}\right) + R_4\right]} \tag{A.6}$$

$$Y_{21b} = 0 \tag{A.7}$$

and

$$Y_{22_b} = G_{b2} + g_{b2}\left(\frac{1}{\beta_2 g_{b2} D + 1}\right) = G_{b2} + g_{b2} C \qquad (A.8)$$

where $D =$

$$\frac{G_4 G_5 + g_{e2}\frac{g_{e3}}{(1-\alpha_3)}(G_4+G_5) + G_4 G_{B3} g_{e2} + (G_5 + g_{e3})(G_4 + g_{e2}) G_{B3}}{g_{e2}\left[\frac{G_4 G_5 g_{e3}}{(1-\alpha_3)} + G_{B3} G_4 (G_5 + g_{e3})\right]} \qquad (A.9)$$

In these expressions, $g_m = \frac{1}{r_m}$ and $G_n = \frac{1}{R_n}$ where m and n are subscripts of the elements shown in Figures 2, 3, 4. Also $R_{e1} + r_{e1} = R_{E1} = \frac{1}{G_{E1}}$ and $R_{b3} + r_{b3} = R_{B3} = \frac{1}{G_{B3}}$.

And the following assumptions have been made:

$$g_{e3} \cong 0$$

$$\beta_2 + 1 \cong \beta_2$$

$$(\beta_2+1) g_{b2} + g_{e2} + g'_{e2} \cong \beta_2 g_{b2} + g_{e2}$$

$$g_{e2} + G_4 + G_{E2} \cong g_{e2} + G_4$$

Typical element values are substituted in these expressions for s.c. admittances of both N_a and N_b and it is found reasonable to make the following approximations for the y parameters of the total network, in addition to the above.

$$Y_{11} = Y_{11a} + Y_{11b} = G_{b1} + G_{L3} + \frac{g_{b1} G_{E1}}{\beta_1 g_{b1} + G_{E1}} \qquad (A.10)$$

$$Y_{12} = Y_{12a} + Y_{12b} \qquad (A.11)$$

$$\simeq \frac{-\alpha_3}{\left(r_{e2} + \frac{r_{b2}}{\beta_2}\right)\left[\left(\frac{R_{b3}}{\beta_3} + \frac{r_{b3}}{\beta_3} + r_{e3}\right)\frac{1}{R_5} + 1\right] + \left[\left(1 + \frac{R_4}{R_5}\right)\left(\frac{R_{b3}}{\beta_3} + \frac{r_{b3}}{\beta_3} + r_{e3}\right) + R_4\right]} \qquad (A.12)$$

$$y_{21} = y_{21a} + y_{21b} \cong \frac{\beta_1 g_{b1} G_{E1}}{\beta_1 g_{b1} + G_{E1}} \qquad (A.13)$$

$$y_{22} = y_{22a} + y_{22b} \cong G_{L1} + G_{b2} + g_{b2} C \qquad (A.14)$$

where $C = \left(\frac{1}{\beta_2 g_{b2} D + 1}\right)$ and D is given by Equation (A.9). The particular choice of a CE stage in N_a and the CC-CB sequence in N_b has been made to minimize loading effects of the feedback circuit and to unilateralize the feedback path so that the assumptions $y_{12a} \cong 0$ and $y_{21b} \cong 0$ made in Equations (A.12) and (A.13) are physically valid.

In order to obtain some useful design equations, we observe that $|y| = y_{11}y_{22} - y_{12}y_{21} \cong -y_{12}y_{21}$ for the gyrator circuit so that

$$z_{12} \cong \frac{-y_{12}}{|y|} \cong \frac{1}{y_{21}} = R_{E1} + \frac{r_{b1}}{\beta_1} = R_{B1} + r_{e1} + \frac{r_{b1}}{\beta_1} \qquad (A.15)$$

$$z_{21} \cong \frac{-y_{21}}{|y|} \cong \frac{1}{y_{12}} \qquad (A.16)$$

$$= -\frac{1}{\alpha_3}\left\{\left[\left(\frac{R_{b3}}{\beta_3} + \frac{r_{b3}}{\beta_3} + r_{e3}\right)\frac{1}{R_5} + 1\right]\left[r_{e2} + \frac{r_{b2}}{\beta_2}\right] + \left[R_4 + \left(1 + \frac{R_4}{R_5}\right)\left(\frac{R_{b3}}{\beta_3} + \frac{r_{b3}}{\beta_3} + r_{e3}\right)\right]\right\} \qquad (A.17)$$

$$z_{11} = \frac{y_{22}}{|y|} \cong \frac{y_{22}}{-y_{12}y_{21}} = -z_{12}z_{21}y_{22} \qquad (A.18)$$

$$= -z_{12}z_{21}\left(G_{L1} + G_{b2} + g_{b2}C\right) \qquad (A.19)$$

$$z_{22} = \frac{y_{11}}{|y|} \cong \frac{y_{11}}{-y_{12}y_{21}} = -z_{12}z_{21}y_{11} \qquad (A.20)$$

$$= -z_{12}z_{21}\left(G_{b1} + G_{L3} + \frac{g_{b1}G_{E1}}{\beta_1 g_{b1} + G_{E1}}\right) \qquad (A.21)$$

$$= -z_{12}z_{21}\left(G_{b1} + G_{L3}\right) - \frac{z_{21}}{\beta_1} \qquad (A.22)$$

CONCLUDING REMARKS

An RC-gyrator filter,employing this gyrator circuit,has been designed in such a way that the sensitivity of its natural frequencies due to variation in the gyration resistance R_0 is minimized.(8) The experimental data is in very good agreement with theoretical frequency response up to an extended limit of 500 Kc.

ACKNOWLEDGMENT

The author gratefully acknowledges the support for the above research, obtained from the Graduate College, University of Minnesota and the National Science Foundation under its Grant No. GP-3017 and is very much indebted to Mr. R. Srinivasagopolan who did all the experimental work reported here.

REFERENCES

1. B. D. H. Tellegan, "The Gyrator, a New Electric Network Element," Phillips Res. Rep., Vol. 3, No. 2 (April, 1948), pp. 81-101.
2. C. L. Hogan, "The Microwave Gyrator," B. S. T. J., Vol. 31, No. 1 (January, 1952), pp. 1-31.
3. J. H. Silverman, D. R. Curran and J. D. Schoeffler, "Passive Electromechanical Gyrators and Isolators," Proc. N. E. C., Vol. 17 (October, 1961), pp. 521-529.
4. M. Onoe and M. Sawabe, "A Piezoelectric-Piezomagnetic Gyrator," Proc. I. R. E., Vol. 50, No. 9 (September, 1962), pp. 1967-1972.
5. G. E. Sharpe, "Ideal Active Elements," J. I. E. E., Vol. 3 (New Series) (July, 1957), p. 431.
6. B. P. Bogert, "Some Gyrator and Impedance Invertor Circuits," Proc. I. R. E., Vol. 43, No. 7 (July, 1955), pp. 793-796.
7. M. S. Ghausi and F. D. McCarthy, "A Realization of Transistor Gyrators," Proc. N. E. C., Vol. 19 (October, 1963), pp. 396-406.
8. D.A.Calahan," Sensitivity Minimization in Active RC Synthesis", Trans.IRE,Vol. CT-9 (March 1962), pp.38-42.

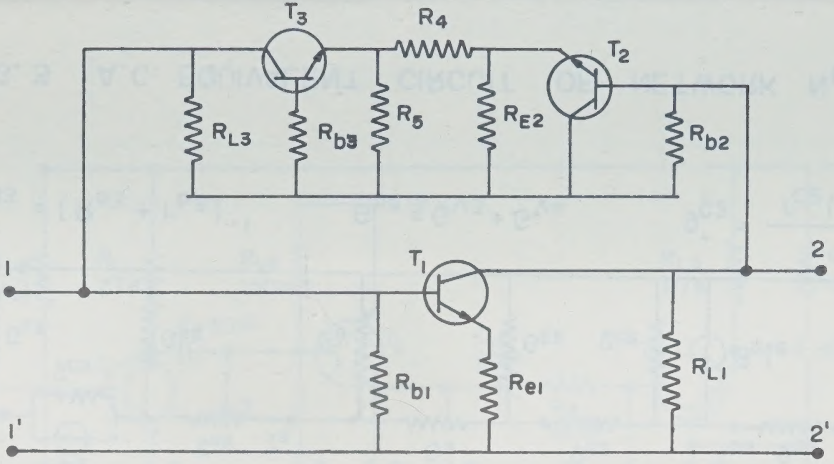

T3
R4
T2
RL3
Rb3
R5
RE2
Rb2
T1
1
2
1'
2'
Rb1
Re1
RL1

FIG. 1 A.C. EQUIVALENT CIRCUIT OF THE GYRATOR

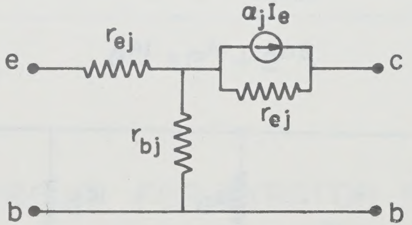
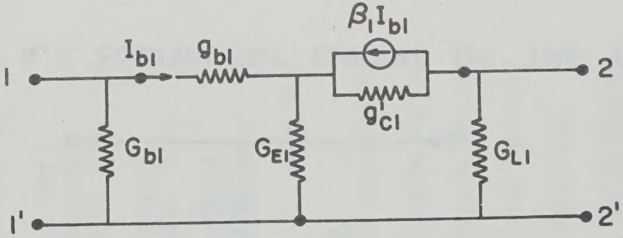
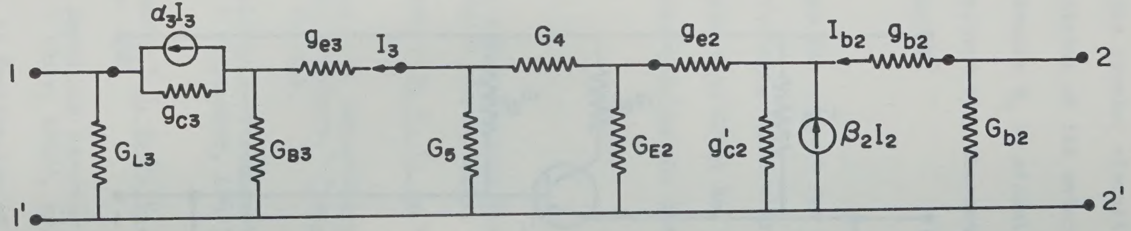

$\alpha_j I_e$
r_{ej}
e
c
r_{ej}
r_{bj}
b
b

FIG. 1.b A.C. EQUIVALENT CIRCUIT OF THE TRANSISTORS


1
1'
2
2'
I_{b1}
g_{b1}
$\beta_1 I_{b1}$
g'_{C1}
G_{b1}
G_{E1}
G_{L1}

$G_{E1} = (R_{e1} + r_{e1})^{-1}$ $\qquad G_{b1} = G_{V1} + G_{V2}$ $\qquad g'_{C1} = \dfrac{1}{r_{C1}(1-\alpha_1)}$

FIG. 2 A. C. EQUIVALENT CIRCUIT OF NETWORK N_a


1
1'
2
2'
$\alpha_3 I_3$
g_{C3}
g_{e3}
I_3
G_4
g_{e2}
I_{b2}
g_{b2}
G_{L3}
G_{B3}
G_5
G_{E2}
g'_{C2}
$\beta_2 I_2$
G_{b2}

$G_{B3} = (R_{b3} + r_{b3})^{-1}$ $\qquad G_{b2} = G_{V3} + G_{V4}$ $\qquad g'_{C2} = \dfrac{1}{r_{C2}(1-\alpha_2)}$

FIG. 3 A.C. EQUIVALENT CIRCUIT OF NETWORK N_b

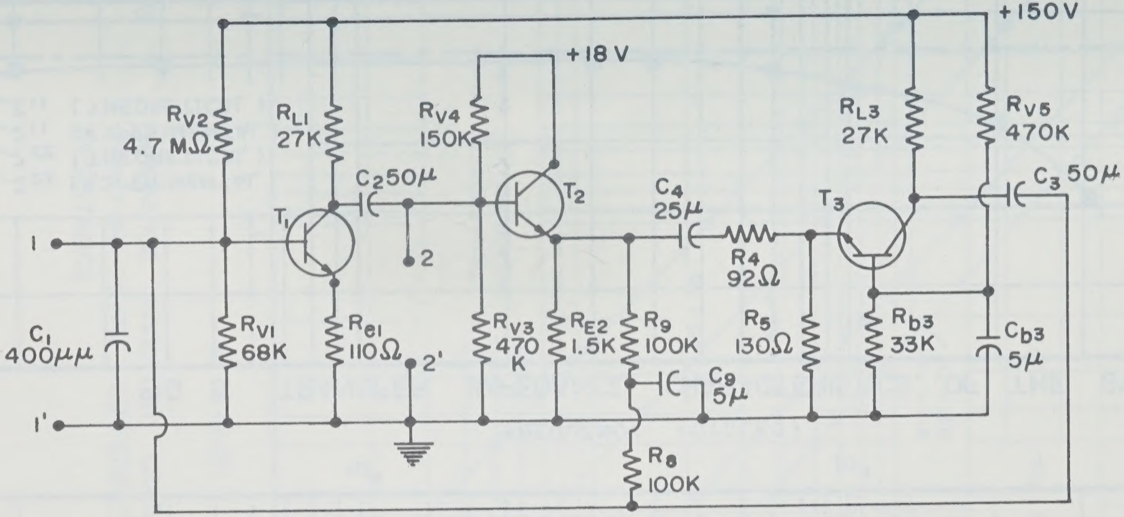

+150V
+18 V
R_{V2} 4.7 MΩ
R_{L1} 27K
C_2 50μ
R_{V4} 150K
T_2
C_4 25μ
R_4 92Ω
T_3
R_{L3} 27K
R_{V5} 470K
C_3 50μ
T_1
1
2
C_1 400μμ
R_{V1} 68K
R_{e1} 110Ω
2'
R_{V3} 470 K
R_{E2} 1.5K
R_9 100K
C_9 5μ
R_5 130Ω
R_{b3} 33K
C_{b3} 5μ
1'
R_8 100K

FIG. 4 ACTUAL CIRCUIT FOR GYRATOR REALIZATION
(ELEMENT VALUES ARE FOR REALIZING $R_0 = 118\Omega$)

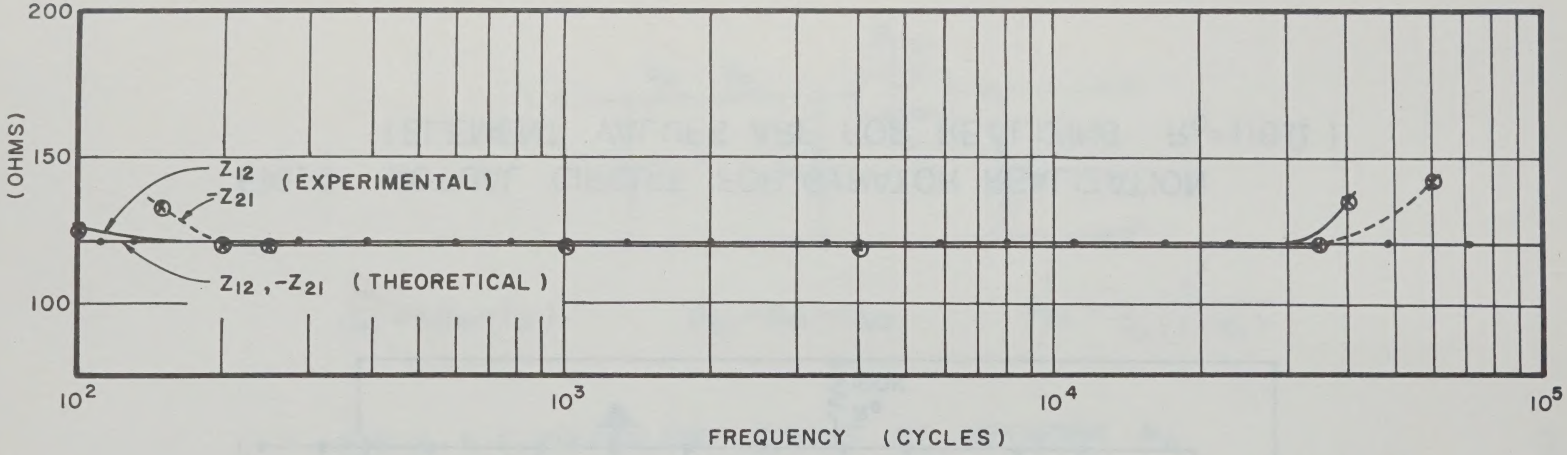


FIG. 5 TRANSFER IMPEDANCE CHARACTERISTICS OF THE GYRATOR

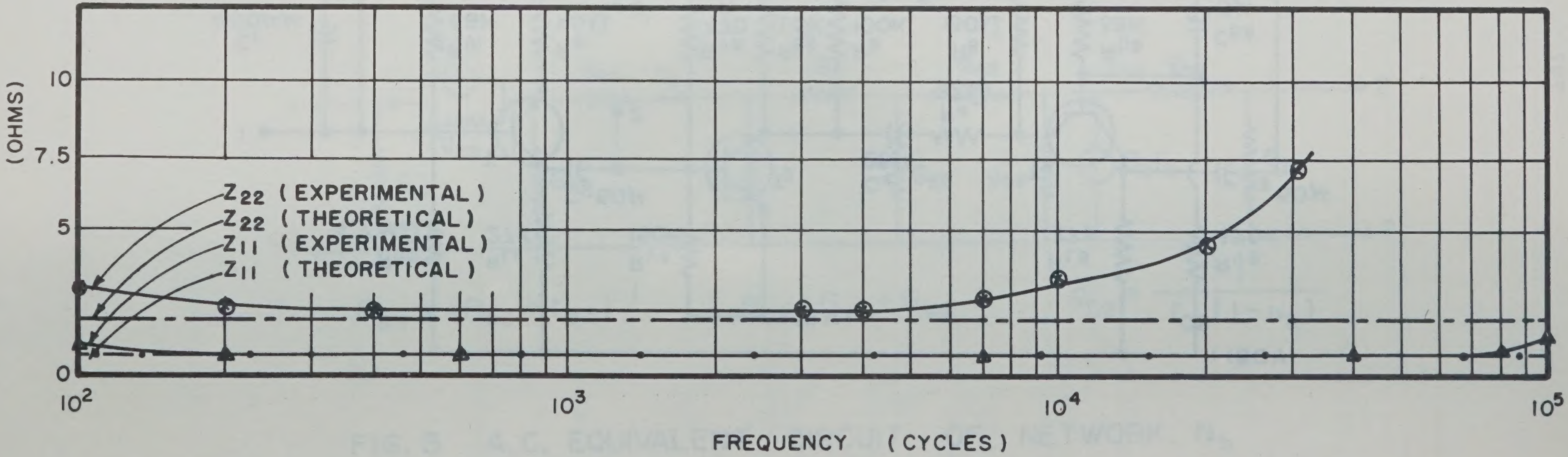


FIG. 6 DRIVING POINT IMPEDANCE
CHARACTERISTICS OF THE GYRATOR

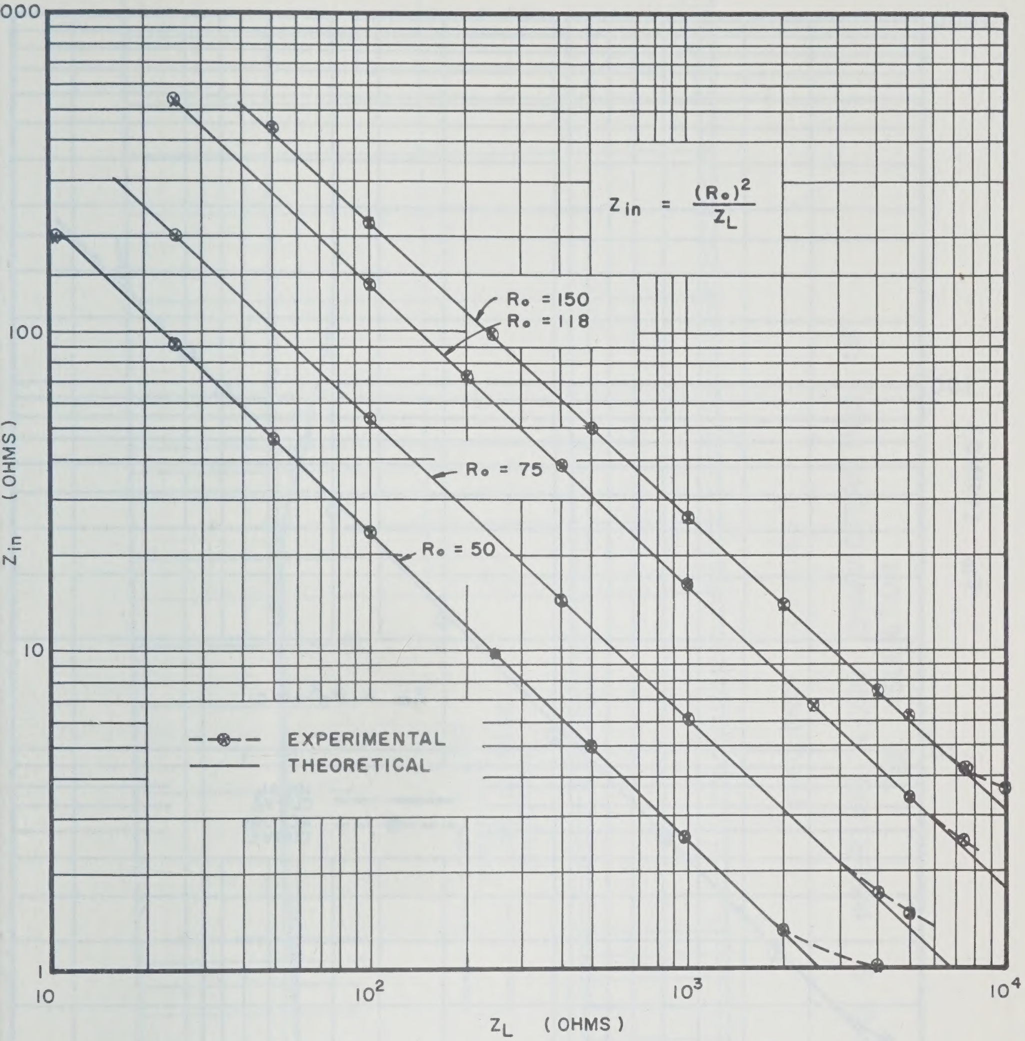


FIG. 7 IMPEDANCE INVERTER CHARACTERISTIS OF THE GYRATOR WITH DIFFERENT GYRATION RESISTANCE

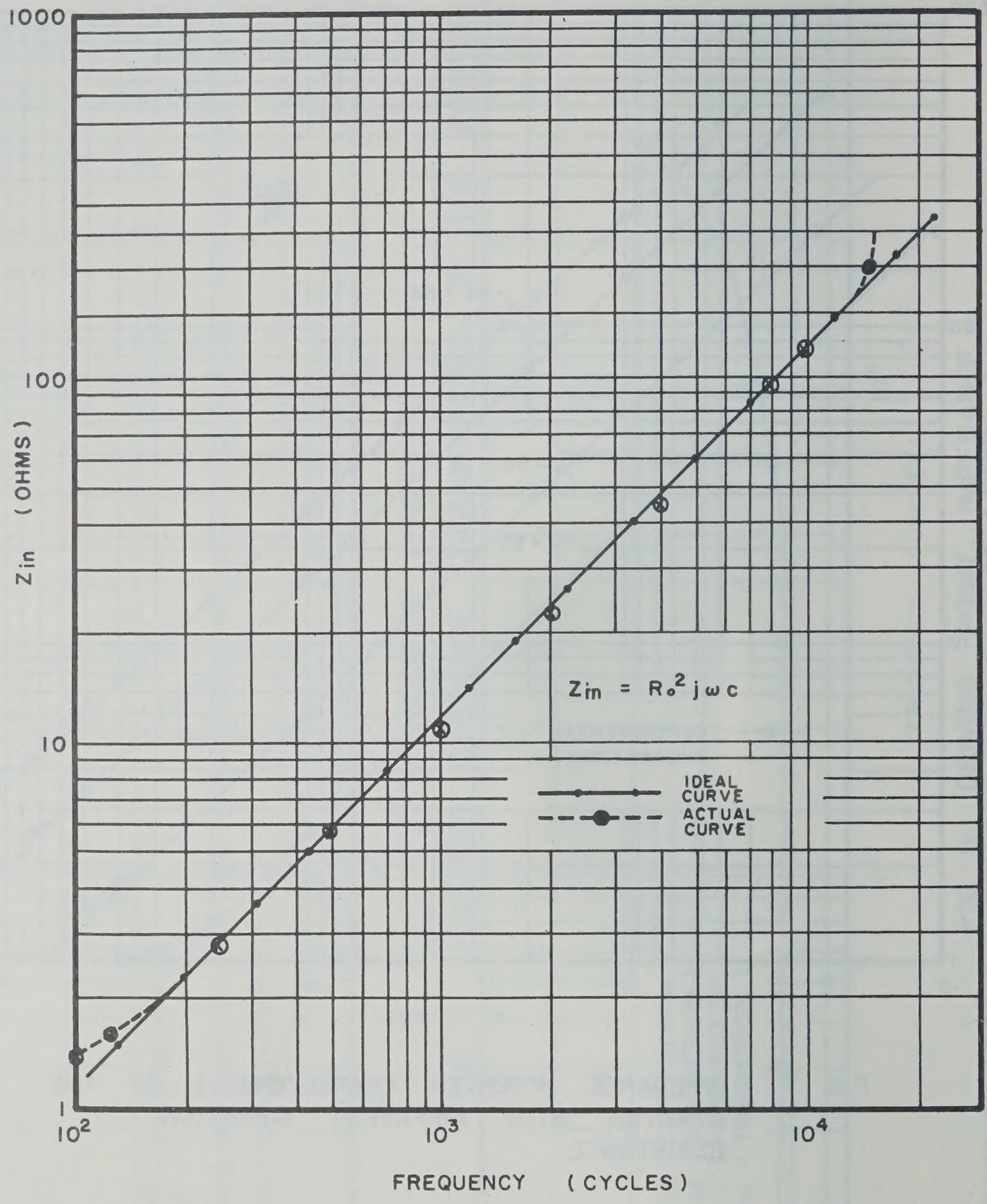


FIG. 8 IMPEDANCE INVERTER CHARACTERISTICS OF THE GYRATOR
(LOAD CAPACITANCE = ·133 μf)

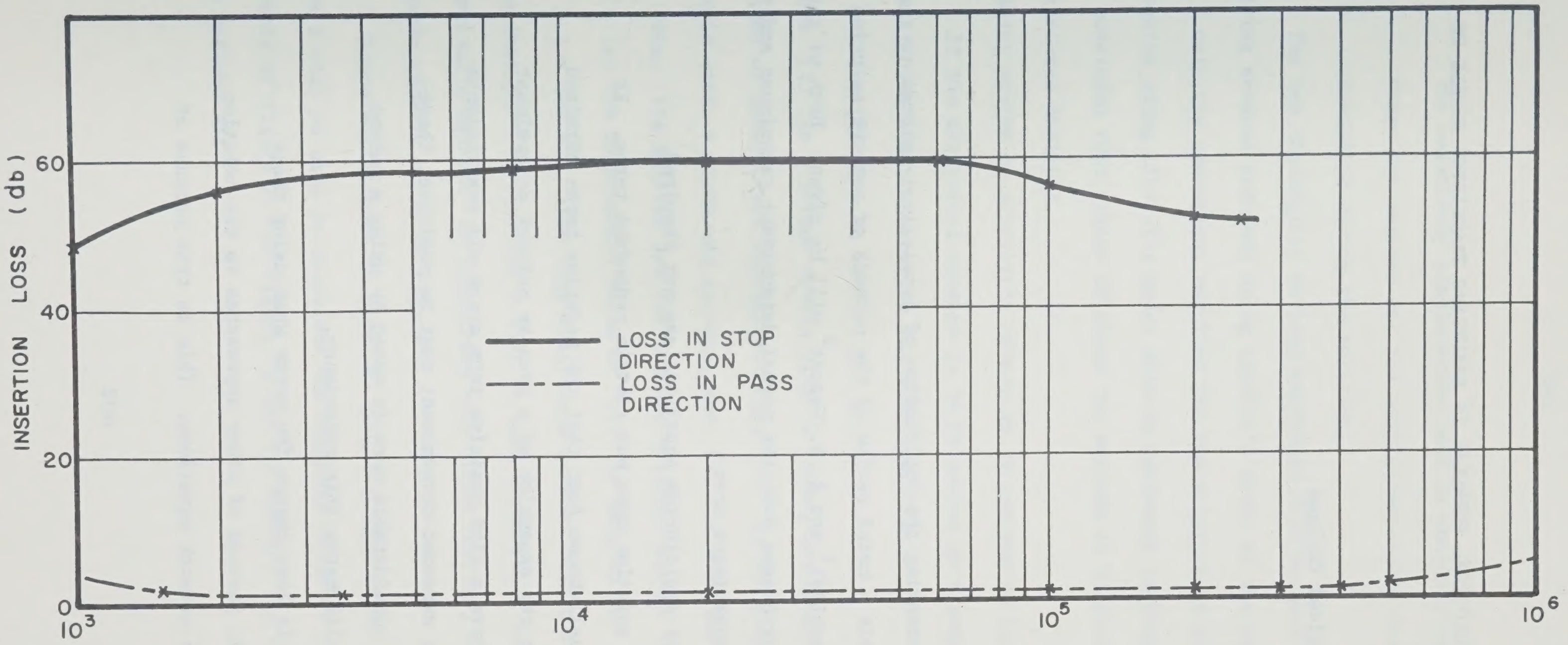


FIG. 9 INSERTION LOSS CHARACTERISTICS OF THE ISOLATOR

THE USE OF TOPOLOGICAL FORMULAS IN ANALYZING TRANSISTOR MODELS OF LINVILL

R. H. Beene
E. D. Merkl
R. H. Seacat
Texas Technological College
Lubbock, Texas

INTRODUCTION

Before presenting the application of topological methods to the transistor models a brief review of the methods of network solution due to W. S. Percival[1] and R. H. Seacat[2] will be given. Both of these methods have advantages over the normal determinant techniques and several of these advantages are:

1. Network equilibrium equations are not required.
2. Their solution does not contain redundant terms. It is a well-known fact that the negative terms obtained during the expansion of a passive network determinant must have a like positive term since all coefficients of the expanded determinant must be positive; therefore, considerable work is saved by using a method that eliminates this redundancy.
3. There is less chance for error when using these methods instead of other approaches to the solution of the network equations. This is true because of

the relatively simple rules used in conjunction with these two methods and the many intermediate checks available during the solution.

The two topological methods mentioned will be used as an aid in solving several problems using Linvill's model of the transistor.

Only the operations required for the solution of transistor circuits using Linvill's model will be reviewed; however, it must not be concluded that either of these two methods is limited to such a restricted analysis.

A BRIEF REVIEW OF PERCIVAL'S METHOD AS IT APPLIES TO PASSIVE NETWORKS

If the electrical network is to be solved on a nodal basis, its equilibrium equations can be put into the form:

$$[I] = [Y][V] \qquad (1)$$

where [I] is a column matrix with N rows, its entries being known values of independent current sources; [Y] is the network admittance matrix with N rows and columns, and [V] is a column matrix with N rows. Its entries are the unknown voltages across the N node pairs.

Let the determinant value of [Y] be denoted by:

$$\Delta = |Y| = \Sigma \text{ tree edge admittance product over all trees.} \qquad (2)$$

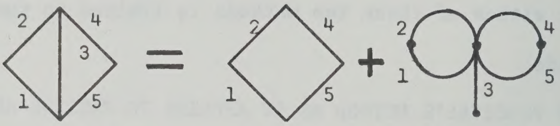
The last part of Equation 2 will be used extensively in applying Percival's method. To illustrate this point further, consider the simple network shown in Figure 1.

Since admittances in parallel add, the simplified graph of Figure 2 will be used to calculate the expanded value of the admittance matrix of Figure 1. In Figure 2, let the numbers 1, 2, 3, 4, and 5 be defined as:

$$\begin{aligned} 1 &= G_1 + SC_1 \\ 2 &= G_2 \\ 3 &= G_3 + SC_3 \\ 4 &= G_4 \\ 5 &= G_5 + SC_5 \end{aligned} \qquad (3)$$

where S is the operator $\frac{d}{dt}$.

Using Percival's "network equations" one can write:



(4)

Equation 4 indicates that the trees of the graph to the left of the equal sign are equal to the trees of the graphs on the right. Note that edge 3 is removed in the first graph on the right of the Equation 4; it is then replaced by a short-circuit, causing edges 1 and 2 to be in parallel in one part of the graph, 4 and 5 to be paralleled in another section of the graph. Since the trees of any separable network are simply the product of the trees of each sub-network, one can write at once, with the use of Equation 4

$$\Delta = 124 + 245 + 451 + 512 + 3(1 + 2)(4 + 5). \qquad (5)$$

From the very definition of a tree it is obvious that all trees, from the same graph, contain exactly the same number of edges and this number must be one less than the number of vertices. This is very useful in checking to see if some careless error has been made while calculating Δ.

If Equations 3 and 5 are combined one has the following results:

$$\begin{aligned}\Delta = {} & C_1C_3C_5S^3 + [C_1C_3G_5 + C_1C_3G_4 + C_3C_5G_2 + C_3C_5G_1 + C_1C_5G_4 \\ & + C_1C_5G_2 + C_1C_5G_3]S^2 + [G_2G_5C_3 + G_2G_4C_3 + G_1G_5C_3 + G_1G_4C_3 \\ & + G_3G_5C_1 + G_3G_4C_1 + G_2G_3C_5 + G_1G_3C_5 + G_1G_2C_5 + G_2G_5C_1 \\ & + G_1G_4C_5 + G_4G_5C_1 + G_2G_4C_5 + G_2G_4C_1]S + [G_1G_2G_4 + G_2G_4G_5 \\ & + G_1G_4G_5 + G_1G_2G_5 + G_1G_3G_4 + G_1G_3G_5 + G_2G_3G_4 + G_2G_3G_5]. \qquad (6)\end{aligned}$$

Note that Equation 6 contains thirty terms. If the equilibrium equations had been written and the determinant of the admittance matrix had been expanded, one would have gotten forty-two terms, twelve of which are redundant.

The network shown in Figure 1 was chosen to demonstrate Percival's method because it is this type of network, with different elements, that is used by Linvill in his model for the transistor. Before moving on to Linvill's model, the Z parameters for this network will be found. Also, voltages at various points will be calculated, assuming the network to be driven at both ports by current sources.

Let

$Z_{11} = \Delta_{11}/\Delta$,

$Z_{12} = Z_{21} = \Delta_{12}/\Delta$ (for passive network),

and

$Z_{22} = \Delta_{22}/\Delta$,

where Δ_{11} is found by first placing a short circuit across the input terminals and then proceeding in exactly the same manner as if the Δ of the remaining graph was desired. Δ_{22} is found through the identical process except the short circuit is placed across the output terminals instead of the input. To calculate Δ_{12}, refer to Figure 3.

A path from the terminal marked 1 and 2 is noted; at the same time another non-touching path is chosen from terminals 1' and 2'. A product is made with all edge admittances used in the paths and with the trees remaining in the graph after all vertices used in the paths are shorted. This process is repeated for all possible combinations of paths that do not touch. The results obtained from the process are then summed over all paths. Next refer to Figure 3b. The same technique as outlined above is followed except the paths are crossed as shown. After summing over all paths, this result is subtracted from the first sum. This difference is Δ_{12}. Note that in all ladder type networks, there can never be any non-touching cross-paths.

Let us place a short across the input terminals of the network of Figure 1. This network has the graph shown in Figure 4a. Figure 4b is the remaining graph after the output terminals are shorted. Therefore, using the graphs of Figure 4, one obtains directly:

$$\Delta_{11} = (G_2 + G_3 + SC_3)(G_4 + G_5 + SC_5) + G_4(G_5 + SC_5) \tag{8}$$

and

$$\Delta_{22} = (G_1 + G_2 + SC_1)(G_3 + G_4 + SC_3) + G_2(G_1 + SC_1). \tag{9}$$

Since there is only one path from input to output meeting the restrictions previously mentioned,

$$\Delta_{12} = G_2G_4. \tag{10}$$

Substituting Equations 8, 9, and 10 into Equations 7 gives the Z parameters of the network.

Next suppose Figure 1 has a current source of magnitude I_1 and I_2 connected at terminals 1-1' and 2-2' respectively. Superposition can be used to determine the voltage at any terminal pair within the network using techniques already learned from Percival. For example, $V_{2-2'}$ is given by the expression

$$V_{2-2'} = Z_{22}I_2 + Z_{12}I_1. \tag{11}$$

If the voltage is desired at some other point, say across terminals A-B as the output of the circuit, driven first at terminals 1-1', and then at terminals 2-2'. The voltage V_{AB} is then

$$V_{AB} = V_{AB1} + V_{AB2} \tag{12}$$

where

$$V_{AB1} = I_1 \frac{(G_2)[G_4 + G_5 + SC_5]}{\Delta}; \tag{13}$$

and

$$V_{AB2} = \frac{I_2(G_4)[G_1 + G_2 + SC_1]}{\Delta}. \tag{14}$$

In the next section these simple operations using Percival's method will be applied to the model of a transistor.

PERCIVAL'S METHOD APPLIED TO LINVILL'S MODEL

In this section the lumped model for a transistor connected in the common-base configuration will be considered using Percival's method. This model for the transistor is due to Linvill.

There are three types of elements used in Linvill's model. These lumped elements are shown in Figure 5. A one-π-section of Linvill's model is shown in Figure 6. It is assumed that the symbols listed in Figure 6 are familiar quantities to the reader; however, a few words of introduction will be sufficient to describe the lumped model. Briefly, the lumped model for a transistor can be shown to be analogous to an electrical network if the following observations are made:

1. The excess density will be treated in this solution as node voltages are used in an electrical network's solution.
2. H_c, H_d, and C will have the same effect as conductance and capacitance in a normal RC network.
3. Current retains its same properties so that one can write

$$I_e = P_{1ex}(H_{c1} + H_d + SC_1) - P_{2ex}H_d \tag{15}$$

which is a normal node equation except excess density is playing the role of node voltage. The excess density is defined by

$$P_{ex} = P_{ON}(e^{qv/KT}-1) \tag{16}$$

where all the terms in Equation 16 are well known quantities.

Lumped models for the transistor, with more than two sections, find application in some situations. It is in these circuits that the topological approach has greatest advantage; however, since the main objective of this paper is to illustrate the topological method in a very simple manner only the one-π-section model will be used. This will in no way take away from the illustration and will save considerable space.

As as example of the use of Pervical's method on a typical transistor problem, suppose that the waveform shown in Figure 7a is applied to the transistor used in a common-base configuration, shown in Figure 6. The other waveforms shown in Figure 7 can be calculated from the lumped model. The storage delay time Γ_D shown in Figure 7d will be calculated, using Percival's method to advantage.

The simplified graph for Figure 6 is shown in Figure 8. Using this graph one writes:

$$\Delta = (H_{c1} + SC_1)(H_d + H_{c2} + SC_2) + (H_d)(H_{c2} + SC_2), \tag{17}$$

$$\Delta_{11} = H_d + H_{c2} + SC_2, \tag{18}$$

$$\Delta_{12} = H_d, \tag{19}$$

and

$$\Delta_{22} = H_d + H_{c1} + SC_1. \tag{20}$$

Since the excess density plays the same role as node voltages in normal passive electrical networks, one may write using Percival's method:

$$P_{1ex(0)} = \frac{E_1}{R_e}\left[\frac{H_d + H_{c2}}{H_{c1}H_d + H_{c1}H_{c2} + H_dH_{c2}}\right] - \frac{E_c}{R_c}\left[\frac{H_d}{H_{c1}H_d + H_{c1}H_{c2} + H_dH_{c2}}\right] \quad (21)$$

$$P_{2ex(0)} = \frac{E_1}{R_e}\left[\frac{H_d}{H_{c1}H_d + H_{c1}H_{c2} + H_dH_{c2}}\right] - \frac{E_c}{R_c}\left[\frac{H_d + H_{c1}}{H_{c1}H_d + H_{c1}H_{c2} + H_dH_{c2}}\right] \quad (22)$$

Equations 21 and 22 were obtained using Equation 17 through 20 and the fact that superposition applies. This is the initial values of the excess densities $P_{1ex(0)}$ and $P_{2ex(0)}$. Next define quantities α_N and α_I, the normal and inverse current gain respectively, as

$$\alpha_N = \frac{H_d}{H_d + H_{c1}} \text{ and } \alpha_I = \frac{H_d}{H_d + H_{c2}} . \quad (23)$$

If $I_c \cong - \frac{E_c}{R_c}$, $I_{E1} \cong \frac{E_1}{R_e}$, and the results of Equation 23 are substituted into Equations 21 and 22, the new expressions are:

$$P_{1ex(0)} = \frac{I_{E1}\alpha_N + I_c\alpha_N\alpha_I}{(1 - \alpha_N\alpha_I)H_d} \quad (24)$$

$$P_{2ex(0)} = \frac{I_c\alpha_I + I_E\alpha_N\alpha_I}{(1 - - \alpha_N\alpha_I)H_d}. \quad (25)$$

The expression for Δ, given in Equation 17, is set equal to zero. After some algebra one obtains the characteristic equation

$$S^2 + S\left[\frac{H_{c2} + H_d}{C_2} + \frac{H_{c1} + H_d}{C_1}\right] + \left[\left(\frac{H_{c1} + H_d}{C_1}\right)\left(\frac{H_{c2} + H_d}{C_2}\right) - \frac{H_d(H_{c1} + H_d)H_d(H_{c2} + H_d)}{C_1(H_{c1} + H_d)(C_2)(H_{c2} + H_d)}\right] = 0. \quad (26)$$

Setting $\omega_{\alpha N} = \frac{H_{c1} + H_d}{C_1}$ and $\omega_{\alpha I} = \frac{H_{c2} + H_d}{C_2}$ and making this substitution into Equation 26 gives:

$$S^2 + S[\omega_{\alpha I} + \omega_{\alpha N}] + \omega_{\alpha I}\omega_{\alpha N}[1 - \alpha_I\alpha_N] = 0. \quad (27)$$

For most transistors $1 - \alpha_I\alpha_N$ is such that the two roots are approximated by

$$S_1 \cong - [\omega_{\alpha I} + \omega_{\alpha N}] \quad (28)$$

$$S_2 \cong \frac{-[\omega_{\alpha I}\omega_{\alpha N}][1 - \alpha_I\alpha_N]}{\omega_{\alpha I} + \omega_{\alpha N}}. \quad (29)$$

If the roots are separated greatly as implied above, then a single time constant approximation can be made to give

$$P_{2ex} \cong \frac{I_c\alpha_I + I_{E2}\alpha_N\alpha_I}{G_d(1-\alpha_N\alpha_I)} + \left[\frac{I_c\alpha_I + I_{E1}\alpha_N\alpha_I}{G_d(1-\alpha_N\alpha_I)} - \frac{I_c\alpha_I + I_{E2}\alpha_N\alpha_I}{G_d(1-\alpha_N\alpha_I)}\right]e^{S_2 t}, \qquad (30)$$

where $I_{E2} \cong - E_2/R_e$. Setting Equation 30 equal to zero and solving for t gives the storage delay time

$$\tau_D \cong \frac{\omega_{\alpha I} + \omega_{\alpha N}}{\omega_{\alpha I}\omega_{\alpha N}(1-\alpha_N\alpha_I)} \ln\left[\frac{I_{E2} - I_{E1}}{I_c/_{\alpha N} + I_{E2}}\right]. \qquad (31)$$

Topological methods were used to advantage in the foregoing calculations as listed below:

1. To find the characteristic equation for the network or lumped-model.
2. To write by inspection the equations controlling the excess densities.
3. To write the expression for the normal and inverted current gain.

A BRIEF REVIEW OF THE DUAL OF FEUSSNER'S METHOD AS APPLIED TO PASSIVE NETWORKS

The equilibrium equations of an electrical network comprised of m-independent nodes and a single source, can be represented by the matrix equation:

$$I] = [G][V] \qquad (32)$$

where [I] is a column matrix of m elements of which only one element is non-zero, [G] is a m x m matrix of admittances and [V] is a column matrix of m node voltages. From Cramer's expansion the voltage response to a current excitation is

$$V_k = \frac{\Delta_{ik} I_i}{\Delta}. \qquad (33)$$

Further, Δ may be expressed in the form

$$\Delta = G_L \Delta_L + \Delta_{L'} \qquad (34)$$

where Δ_L is the determinant derived from the principal determinant by collapsing the row and column containing G_L and $\Delta_{L'}$ is the principal determinant with G_L being set equal to zero.

By finding a topological correspondence between the above expression of Δ and that of the actual network, one may develop the table of correspondences shown in Table 1.[2] Thus one needs only to extremize the elements in a circuit in such a manner to obtain any of the topological circuits in Table 1. The circuit or network remaining after an element is extremized, i.e., either set equal to zero or infinity, is called a residue circuit denoted by the subscripts of the Δ's. The element which is set equal to the extreme value of zero is denoted by the primed subscript.

Applying this technique to the circuit of Figure 1, one finds Δ as follows:

Setting SC_1, SC_2, and SC_3 equal to infinity as shown in Figure 9a, the residue value is easily seen to be 1. Thus, this contribution to Δ is $S^3C_1C_2C_3$. Next, considering the SC_M set to infinity in pairs, and otherwise to zero, the residue networks are 9b, 9c, and 9d, respectively for $S^2C_1C_3$, $S^2C_1C_5$ and $S^2C_3C_5$ with residue values of $(G_4 + G_5)$, $(G_2 + G_3 + G_4)$ and $(G_1 + G_2)$, respectively. Thus the coefficient of S^2 is $C_1C_3(G_4 + G_5)$ $+ C_1C_5(G_2 + G_3 + G_4) + C_3C_5(G_1 + G_2)$.

For the coefficients of S the SC's are set equal to infinity one at a time and equal to zero otherwise. If SC_1 is infinite the residue network is 9e, with a value of $G_2(G_4 + G_5) + G_3(G_4 + G_5)$ $+ G_4G_5$, with SC_2 infinite the residue network is 9f with a value of $G_4(G_2 + G_1) + G_3(G_1 + G_2) + G_1G_2$, and with SC_3 infinite the residue value of its network shown in 9g is $(G_1 + G_2)(G_4 + G_5)$ thus the coefficient of S is the sum of these terms. Finally if all SC's are set equal to zero one has the residue network in 9h, with a value $G_3(G_1 + G_2)(G_4 + G_5) + G_1G_2(G_4 + G_5) + G_4G_5(G_1 + G_2)$. Summing each of the terms one has the same expression as that of Equation 6.

In order to calculate the numerator of the voltage expression, i.e., Δ_{ik}, path voltages,[2] or the voltages which will exist across a graph when there is no complete circuit (with all of the nodes connected), and with the minimum number of elements removed, are used. If the path voltage across a given path f is given by V_f, it is related to I_i/Δ by the ratio β_f; therefore,

$$V_f = \frac{\beta_f I_i}{\Delta}. \tag{35}$$

Since the voltage across a branch k, is the response to all currents in the i branches the total response is

$$V_k = \frac{\Sigma_f \beta_f I_i}{\Delta} \tag{36}$$

and as previously shown

$$V_k = \frac{\Delta_{ik} I_i}{\Delta}\ . \tag{37}$$

It is concluded that $\Delta_{ik} = \Sigma_f \beta_f$. Thus in finding the sum of the residues of all voltage paths comprising any V_k, Δ_{ik} is computed.

As an example, consider the conductance portion of Figure 1, with a current source, connected to terminals 1-1'. The only three possible path voltage networks with their residues β_1, β_2, and β_3, respectively are shown in Figures 10a, 10b, and 10c. Thus the voltage across G_1 is expressed as a function of I, in the form:

$$V_1 = \frac{(\beta_1 + \beta_2 + \beta_3)}{\Delta} I \tag{38}$$

or

$$V_1 = \frac{[G_3(G_4 + G_5) + G_4G_5 + G_2G_4 + G_2G_5]}{\Delta} I. \tag{39}$$

The expression for the voltage across G_2 would be

$$V_2 = \frac{\beta_1 I}{\Delta} \tag{40}$$

or

$$V_2 = \frac{G_3(G_4 + G_5) + G_4G_5}{\Delta}. \tag{41}$$

FEUSSNER'S DUAL APPLIED TO LINVILL'S MODEL

To find the characteristic equation of the model of Figure 6, it is only necessary to calculate Δ. Noting that

$$\Delta = S^2C_1C_2\Delta_{12} + S(C_1\Delta_{12'} + C_2\Delta_{1'2}) + \Delta_{1'2'} \tag{42}$$

Setting both C_1 and C_2 equal to infinity, the residue circuit becomes that shown in Figure 11a, having a residue value of 1. Next let C_1 be infinite and C_2 be zero, the residue circuit is that seen in Figure 11b whose value is $(H_{c2} + H_d)$. With the roles of C_1 and C_2 reversed, the residue circuit is that of Figure 11c, with a residue value of $(H_d + H_{c2})$. Setting both C_1 and C_2 equal to zero, the residue circuit seen in Figure 11d results. Then a value of $H_d(H_{c1} + H_{c2}) + H_{c1}H_{c2}$ will result. Summing the respective coefficients of S, Δ is given by

$$S^2C_1C_2 + S(C_1[H_{c2} + H_d] + C_2[H_{c1} + H_d]) + H_d(H_{c1} + H_{c2}) + H_{c1}H_{c2} \tag{43}$$

which is the expression given in Equation 17.

By determining P_{1ex} at frequency zero, the storances are deleted and P_{1ex} and P_{2ex} are treated analogously as node voltages. The path voltage graphs and their respective β's are shown in Figure 12. One should note each current source is treated individually for the β's, i.e., superposition applies. Obviously

$$\Delta = H_d(H_{c1} + H_{c2}) + H_{c1}H_{c2} \tag{44}$$

it is then a simple task to find

$$P_{lex} = \frac{\beta_1 + \beta_2}{\Delta} \frac{E_1}{R_e} - \frac{\beta_3}{\Delta} \frac{E_c}{R_c} \tag{45}$$

or

$$P_{lex} = \frac{H_d + H_{c2}}{\Delta} \frac{E_1}{R_e} - \frac{H_d}{\Delta} \frac{E_c}{R_c} \tag{46}$$

which is the same expression shown in Equation 21.

CONCLUSIONS

Topological methods can be used to advantage in solving problems dealing with transistor models. The determination of the characteristic equation of the network was found by Percival's method and the dual to Feussner's method, one having no obvious advantage over the other. However, numerator calculations are, in the opinion of the authors, easier using the dual to Feussner's method. Both of these methods are much faster and less subject to error than the normal matrix method.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Dr. Paul G. Griffith for his advice on the transistor section of this paper.

REFERENCES

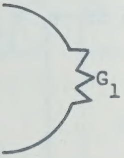
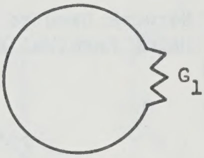
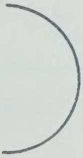
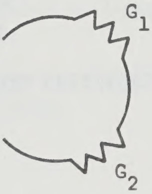
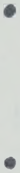
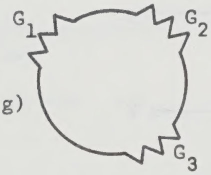
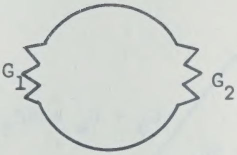
1. W. S. Percival. "Solution of Passive Electrical Networks by Means of Mathematical Trees," Journal of the Institute of Electrical Engineers, London, (Vol. 100), Part III, 1953, Pp. 143-150.

2. R. H. Seacat. "The Dual to Feussner's Method," Paper given 1964 SWIEEECO, Dallas, Texas.

3. J. G. Linvill and J. F. Gibbons. Transistors and Active Circuits, New York: McGraw-Hill Book Co., Inc., 1961.

TABLE 1

Correspondences of Simple Networks

	Topology	Determinant		Topology	Determinant
(a)		$\Delta = G_1$	(e)		$\Delta = 1$
(b)		$\Delta = 1$	(f)		$\Delta = G_1G_2$
(c)		$\Delta = 0$	(g)		$\Delta = G_1(G_2+G_3)+G_2G_3$
(d)		$\Delta = G_1+G_2$			

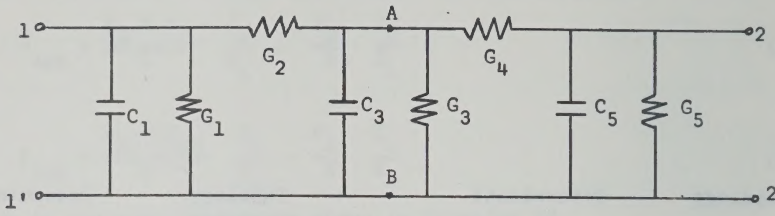

1
A
2
G2
G4
C1
G1
C3
G3
C5
G5
B
1'
2'

Figure 1. Network Used to Illustrate the Calculation of Δ Using Percival's Method.

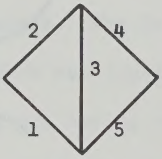
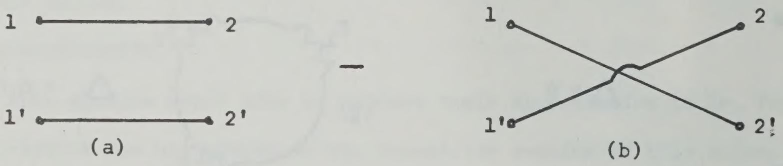
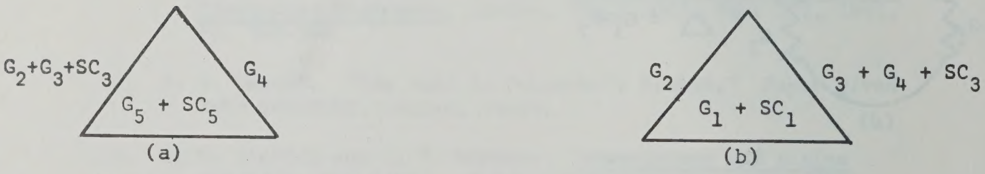


Figure 2. Simplified Graph for Figure 1.



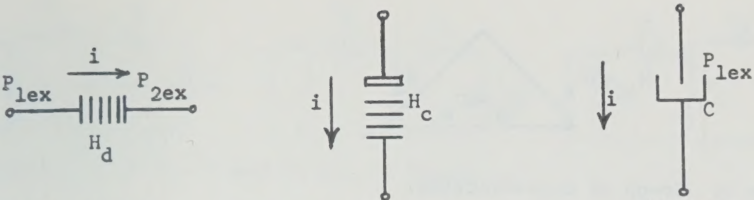
(a) (b)

Figure 3. Paths Used to Calculate Δ_{12}.


G2+G3+SC3
G4
G5 + SC5
G2
G3 + G4 + SC3
G1 + SC1

(a) (b)

Figure 4. (a) Graph With Input Shorted; (b) Graph With Output Shorted.


i
P1ex
P2ex
Hd
i
Hc
i
P1ex
C

Diffusance Combinance Storance

Figure 5. The Three Elements for Linvill's Model.

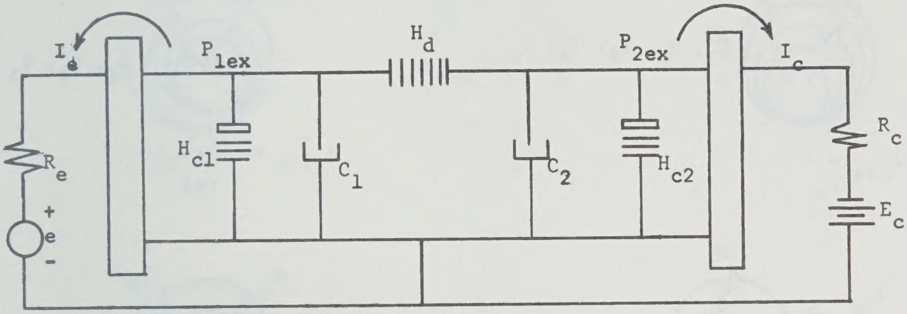

Ie
P1ex
Hd
P2ex
Ic
Re
+
e
-
Hc1
C1
C2
Hc2
Rc
Ec

Figure 6. A One-π-Section Model of a Transistor.

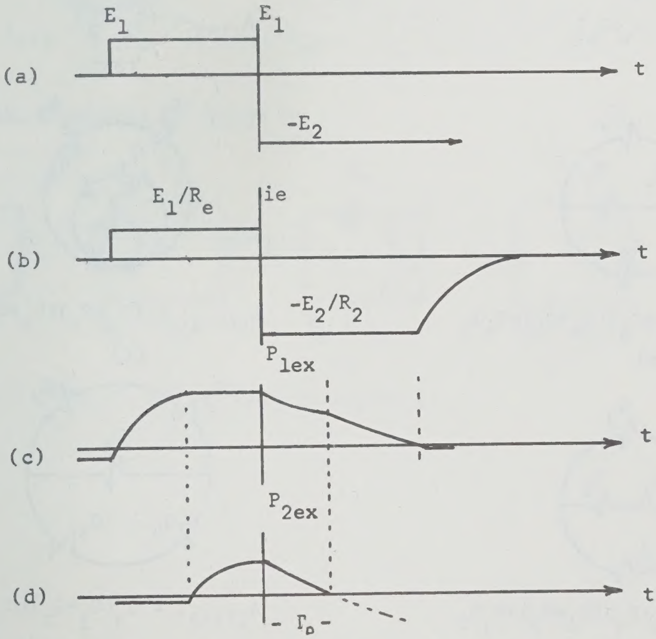


Figure 7. Input to a Transistor and Resulting Transient Responses.

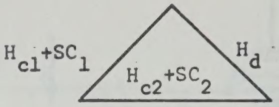
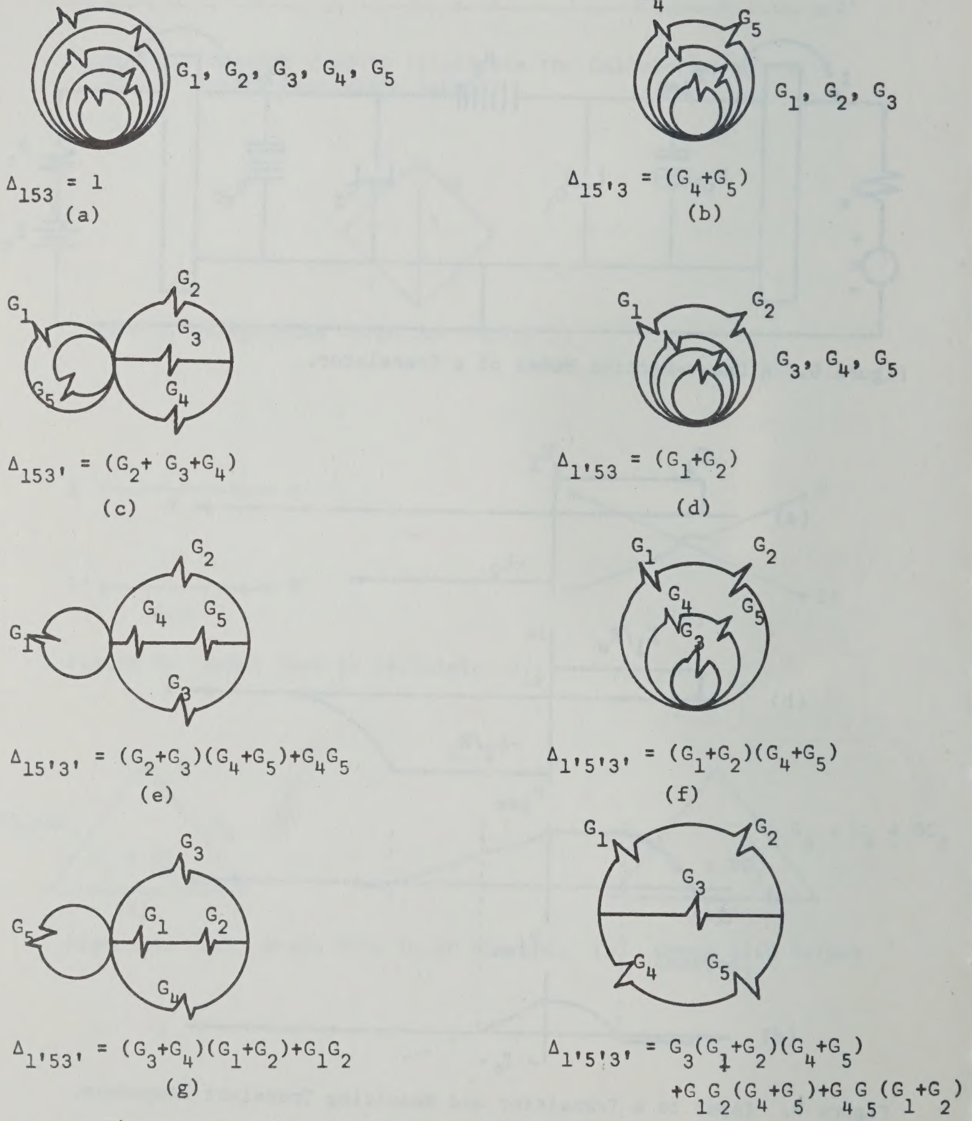

$H_{c1}+SC_1$
$H_{c2}+SC_2$
H_d

Figure 8. Graph of One-π-Section.


G1, G2, G3, G4, G5
G4
G5
G1, G2, G3
G1
G2
G3
G5
G4
G1
G2
G3, G4, G5
G2
G4
G5
G1
G3
G1
G2
G4
G5
G3
G3
G1
G2
G5
G4
G1
G2
G3
G4
G5

$\Delta_{153} = 1$
(a)

$\Delta_{15'3} = (G_4+G_5)$
(b)

$\Delta_{153'} = (G_2+ G_3+G_4)$
(c)

$\Delta_{1'53} = (G_1+G_2)$
(d)

$\Delta_{15'3'} = (G_2+G_3)(G_4+G_5)+G_4G_5$
(e)

$\Delta_{1'5'3'} = (G_1+G_2)(G_4+G_5)$
(f)

$\Delta_{1'53'} = (G_3+G_4)(G_1+G_2)+G_1G_2$
(g)

$\Delta_{1'5'3'} = G_3(G_1+G_2)(G_4+G_5) + G_1G_2(G_4+G_5)+G_4G_5(G_1+G_2)$
(H)

Figure 9. Residue Networks of Figure 1.

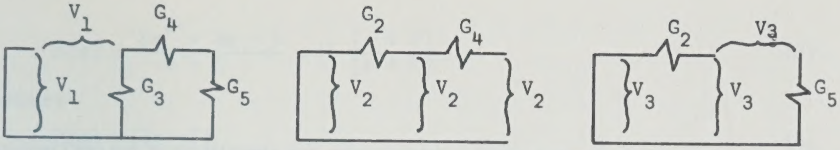
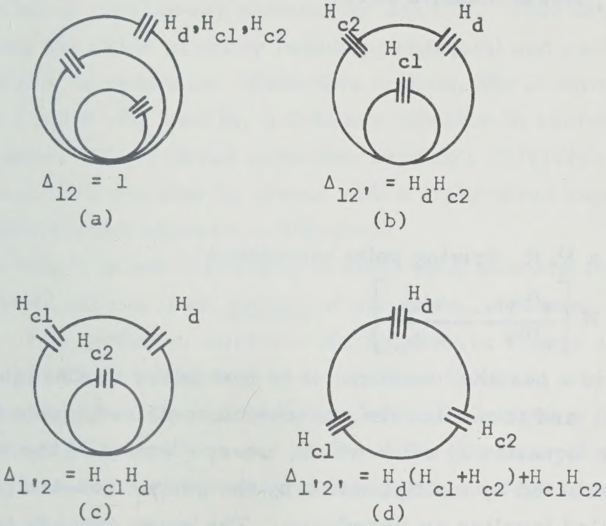

V1
G4
V1
G3
G5
G2
G4
V2
V2
V2
G2
V3
V3
V3
G5

Figure 10. Path Voltage Graphs of Figure 1.


Hd, Hc1, Hc2
Hc2
Hd
Hc1
Hc1
Hd
Hc2
Hd
Hc1
Hc2

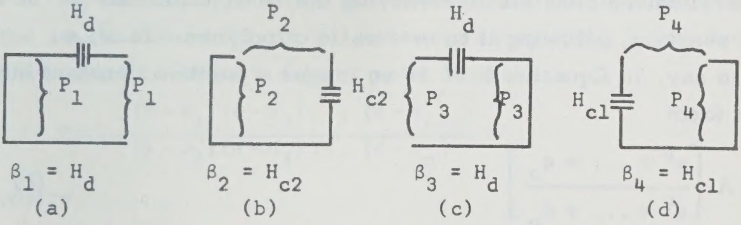
$\Delta_{12} = 1$
(a)

$\Delta_{12'} = H_d H_{c2}$
(b)

$\Delta_{1'2} = H_{c1} H_d$
(c)

$\Delta_{1'2'} = H_d(H_{c1}+H_{c2})+H_{c1}H_{c2}$
(d)

Figure 11. Residues of Figure 6.


Hd
P1
P1
P2
P2
Hc2
Hd
P3
P3
P4
Hc1
P4

$\beta_1 = H_d$
(a)

$\beta_2 = H_{c2}$
(b)

$\beta_3 = H_d$
(c)

$\beta_4 = H_{c1}$
(d)

Figure 12. β's for the DC Form of Figure 6.

GENERALIZED IMPEDANCE LEVELING IN NETWORK SYNTHESIS

Marco Antonio Murray Lasso
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Given a P. R. driving point impedance

$$Z(s) = H\left[\frac{s^n + \ldots + a_o}{s^m + \ldots + b_o}\right] \qquad (1)$$

where H is a positive constant, it is customary to disregard the constant H and to realize the expression contained inside the brackets in Equation 1; after which, every element of the impedance is corrected by multiplication by the proper constant. The process is called leveling an impedance. The same process can be applied to an admittance. In this paper the process is generalized. The generalization consists in removing the restriction that H be a positive constant, allowing it to be a ratio of polynomials of s. That is to say, in Equation 1 H is no longer a positive constant but is of the form

$$H = A\left[\frac{s^p + \ldots + c_o}{s^q + \ldots + d_o}\right] \qquad (2)$$

where $A, c_{p-1}, \ldots, c_o, d_{p-1}, \ldots, d_o$ are real positive constants. The realization of Equation 1 is made as before; and at the end, the impedances are multiplied by the factor H and are replaced by the realization of these products. A trivial example will best illustrate the procedure.

Consider the driving point impedance

$$Z(s) = \frac{2s^2 + 8s + 6}{s + 2} = H \left[\frac{s + 3}{s + 2}\right] \qquad (3)$$

where

$$H(s) = 2(s + 1) . \qquad (4)$$

Just as when H is a positive constant, the quantity in the brackets is realized and the elements are leveled using Equation 4. The realization of $(s+3)/(s+2)$ is given in Figure 1.

Now let us level every element by $2(s+1)$. This is done by multiplying the value of every impedance by $2(s+1)$ and realizing the resultant new impedances. When this is done, the 1-ohm resistor becomes $1 \times 2(s+1)$; that is, a 2-henry inductor in series with a 2-ohm resistor. The 1-farad capacitor becomes $(1/s)2(s+1) = 2 + 2/s$; that is, a 2-ohm resistor in series with a (1/2)-farad capacitor. The leveled circuit appears in Figure 2.

Note that it is not necessary to level each element separately, it is possible also to level groups of elements, obtaining different circuits. For instance, consider the RC tank in Figure 1 as a single impedance. The leveled tank becomes $2(s+1)/(s+2) = 1 + s/(s+2)$. Therefore, another circuit realizing Equation 3 appears in Figure 3.

Numerous applications can be found for this very simple trick. Following are some of the ways of using this concept as an aid in the synthesis of networks.

1. Realization of Two Element Circuits.

An obvious application that stems from the preceding example is the realization of RL or RC networks by splitting it into factors. Given an RL network driving point impedance

$$Z(s) = \frac{(s - s_1)(s - s_3) \ldots (s - s_{n-1})}{(s - s_2)(s - s_4) \ldots (s - s_n)} \qquad (5)$$

where

$$0 \geq s_1 > s_2 \ldots > s_n .$$

The impedance expressed by Equation 5 can be realized as follows: Take as H the product of the poles and zeros enclosed in the rectangle in Figure 4. Realize the rest of the impedance which is RL. When leveling element by element, the L's produce RL impedances and the R's produce RC impedances. Obviously, there are many choices.

2. Criterion for Realizability and Method of Realization.

There are cases where it is possible to tell by inspection if an impedance is realizable by splitting it in two factors and noting that the realization of one of the factors is levelable by the second factor. A class of networks that fall in this category are RLC networks whose poles and zeros lie in the real axis and which can be split in the product of an RL impedance times an RC impedance. The method used is very similar to the one mentioned in Section 1. Given

$$Z = Z_{RC} \cdot Z_{RL} \tag{6}$$

the impedance is always realizable, Reference 1. A very simple proof can be given using generalized impedance leveling: Take $H = Z_{RL}$. Realize Z_{RC} and level each element of Z_{RC} by H. This is always possible since $(1/Cs)(Z_{RL})$ is RC realizable. (See Example 1 in Section 3.) Thus if Z_{RC} is realized in the Foster Form of Figure 5, the realization of Z can be obtained in the form of Figure 6 where the boxes represent RC and RL networks. Here also there exist many possibilities. For instance, the pole-zero pattern of Figure 7 is realizable in the following manner. Take as H the Pole and Zero expression enclosed by the broken line, the rest of the impedance is RC and every element can be leveled by H which is RL.

3. The Leveling Process as a Residue Shifter.

One useful way of thinking of the leveling process is to consider that it controls the position of the residues at the different poles of the impedance. Given an impedance whose residues lie in certain positions, leveling by a factor may relocate the residues in more convenient regions. When the factor is not already a part of the impedance, the process of leveling and releveling (cancelling the first factor) may be used advantageously. The tours that the residues take during the process may be illuminating, since it may give the realization of the impedance. This way of thinking, applied to a well-known instance, is illustrated in Example 1.

Example 1: The RL impedance

$$Z(s) = A \frac{(s - s_1)(s - s_3) \ldots (s - s_{n-1})}{(s - s_2)(s - s_4) \ldots (s - s_n)} \tag{7}$$

where $A > 0 \geq s_1 > s_2 > . . . > s_n$ has all its residues on the negative real axis (the factor, $s-s_n$, may be missing). However, if Z(s) is multiplied by 1/s all the residues are shifted to the positive real axis; this is so because the resulting impedance is RC. Z(s)/s may be decomposed in its partial fraction expansion

$$Z(s)/s = k + k_o/s + k_2/(s-s_2) + . . . + k_n/(s-s_n). \qquad (8)$$

Releveling by s returns us to the original impedance in the form

$$Z(s) = ks + k_o + k_2 s/(s-s_2) + . . . + k_n s/(s-s_n). \qquad (9)$$

It is well known that a very convenient way of calculating the k's of Equation 9 is to make the partial fraction of expansion of Equation 8, (Reference 2). This idea can be generalized and extended to the RLC case.

Express the partial fraction expansion of a P. R. function with complex poles in the form

$$Z(s) = k_\infty s + k + k_1/(s-s_1) + k_1{}^*/(s-s_1) + . . . + k_n/(s-s_n)$$
$$+ k_n{}^*/(s-s_n^*) \qquad (10)$$

If all the residues k_ν, lie in the region A shown in Figure 8, the impedance is easily realized.[3] The realization is made with a series of tanks of the form of Figure 9, plus possibly an inductance and resistance in series to account for k and k_∞ of Equation 10. If the residues do not lie in region A of Figure 8, it is sometimes possible to move them into the region by adding another pole or zero to the impedance. Once the residues are in the region, the altered impedance can be realized by a structure of the form of Figure 9. After the altered impedance has been realized, we try to cancel the added pole or zero by leveling with the proper factor. This is illustrated with Example 2.

Example 2: Consider the impedance

$$Z(s) = \frac{s^2 + 4s + 7}{s^2 + 2s + 2} \qquad (11)$$

which can be written

$$Z(s) = \frac{(s + 2 + j\,3)(s + 2 - j\,3)}{(s + 1 + j)\ \ (s + 1 - j)} = 1 + \frac{1 + 3j/2}{s + 1 + j} + \frac{1 - 3j/2}{s + 1 - j} \qquad (12)$$

The pole-zero pattern of Equation 12 is shown in Figure 10 where k_a and $k_a{}^*$ are the residues at the poles s_a and $s_a{}^*$ respectively; s_b and

s_b* are the zeros. Since the residues lie outside region A, the impedance cannot be realized with a structure as that of Figure 9. However, the addition of a zero on the real axis will move the residue k_a into region A. According to Figure 10

$\alpha = \tan^{-1} 3/2 = 56° \, 19'$, $56° \, 19' - 45° = 11° \, 19'$, $\tan 11° \, 19' = 1/5$. A factor (s + m) in the demoninator in which $0 < m < 6$ will shift k_a into region A. The factor s + 6 will place k_a on the dividing line ℓ. The altered impedance is

$$Z'(s) = \frac{(s^2 + 4s + 7)(s + 6)}{s^2 + 2s + 2} = \frac{s^3 + 10s^2 + 31s + 42}{s^2 + 2s + 2} \qquad (13)$$

A realization of Z'(s) is shown in Figure 11 where the resistance in parallel with the capacitance is missing because the residues are on the dividing line of region A of Figure 10. We should now level the circuit of Figure 11 by $H = 1/(s + 6)$. Note that it is not possible to level every element because the capacitor will not yield a P. R. impedance. If the capacitor had enough resistance in series with it, it could be leveled without trouble. A method of incorporating some resistance in series with the capacitive branch is to replace the T formed inside points a, b, c of Figure 11 by an equivalent Π. The circuit appears in Figure 12, where the resistance in the capacitive branches is sufficient for leveling. The leveled circuit is shown in Figure 13. In the example shown the impedance could have been altered by adding a zero on the real axis with a smaller m which would have brought the residues further into region A and we could have obtained a whole variety of circuits.

4. Circuits with Resistive Behavior at all Frequencies

The leveling process is a convenient way of obtaining circuits that behave resistively at all frequencies. Given a one ohm resistor, level it by a factor s + a to obtain the circuit of Figure 14. Return to the original impedance by releveling element by element by 1/(s + a). The circuit is shown in Figure 15. If, instead, the one-ohm resistor is first leveled by 1/(s + a) and later each element is leveled by s + a, the circuit of Figure 16 is found. To obtain R ohms, level the circuits by R. Note that a has not been specified yet, since it is at the disposal of the designer, and the only limitation is $a > 0$. More complicated networks having resistive input impedance may be obtained by leveling with higher degree polynomials. For instance, both circuits of Figure 17 have an impedance of one ohm. A large family of circuits that behave resistively may be obtained through frequency transformations of these.[4]

COMMENTS ON THE APPLICABILITY OF THE LEVELING PROCESS

Leveling an impedance by a positive constant will never spoil the P. R. character of a function, but generalized leveling may. This means that it is not always possible to level a specified impedance by a given factor and end with a realizable circuit. In example 2, for instance, leveling was impossible until enough resistance was incorporated in the capacitive branches. This condition forces the networks' configurations to be of certain forms, depending on the functions to be realized. The condition that is usually troublesome about P. R. functions is that the real part may not be negative for $s = j\omega$. The closer an impedance comes to violating this condition, the more complicated becomes the structure, and the less the choices of leveling we can make. If an impedance is minimum resistive (on the verge of violating the mentioned condition) the choices are drastically reduced; the factors have to be chosen carefully, and the leveling cannot be done element by element but the whole impedance has to be leveled simultaneously. The minimum biquadratic impedance will be treated in the paper, while the general RLC impedance is discussed in the Appendix

5. Realization of a Biquadratic Minimum Function through Generalized Leveling

Given a P. R., minimum, biquadratic impedance reduced to the form

$$Z(s) = \frac{s^2 + as + b}{s^2 + cs + d}, \qquad (14)$$

where $a, b, c, d > 0$. The real part for $s = j\omega$ has zeros whose position is given by

$$\omega_{1,2} = \pm\sqrt{\frac{b + d - ac}{2}} = \pm m. \qquad (15)$$

The zeros at m are double, consistent with the fact that $Z(s)$ is P. R. and minimum. Because the impedance is minimum resistive, we must be careful with the real part for $s = j\omega$. We will try to reduce the algebraic expression by extracting impedances with poles or zeros on the imaginary axis; this will not affect the real part of the impedance. To accomplish this, let us level the impedance by 1/(s + h) and then relevel the whole structure by s + h. The impedance now takes the form

$$Z(s) = \frac{s^3 + (a+h)s^2 + (b+ha)s + bh}{s^3 + (c+h)s^2 + (d+hc)s + dh} = Z_1 + Z_2, \qquad (16)$$

Split Equation 16 into two parts, one with zeros on the jω axis at $\omega = \pm m$ plus a remainder with a zero at the origin

$$Z(s) = \frac{(bh/m^2)(s^2 + m^2)}{s^3 + (c+h)s^2 + (d+hc)s + dh} + \frac{s^3 + (a + h - bh/m^2)s^2 + (b+ha)s}{s^3 + (c+h)s^2 + (d+hc)s + dh} \qquad (17)$$

If h is determined in such a way that the real part of Z_1 for $s = j\omega$ has double zeros at $\pm m$, both Z_1 and Z_2 (the first and second terms on the right side of Equation 17 respectively), are realizable as proven below. This requires

$$m^2 = \frac{b + d - ac}{2} = dh/(c+h) \qquad (18)$$

or,

$$h = \frac{b + d - ac}{d - b + ac} \cdot c \,. \qquad (19)$$

It may happen that h turns out to be negative, in which case we realize $Z' = 1/Z$ and then find the dual of Z' which will realize Z. The switch is equivalent to interchanging a and c, and b and d. The numerator of Equation 19 is always positive, since $\omega_{1,2}$ in Equation 15 is real. Thus, if h were negative, $d - b + ac < 0$ which implies

$$d < b \qquad (20)$$

Therefore, the interchange of d and b will alter the sign of h.

In the case h is undefined because the denominator of Equation 19 vanishes, the indetermination will disappear for a 1/Z and furthermore h will be positive since $d - b + ac = 0$ implies $d - b < 0$, which is the same as Expression 20. Having guaranteed $h > 0$, we now prove that both Z_1 and Z_2 of Equation 17 are P. R. The denominator of both is Hurwitz since $h > 0$ and the denominator of Equation 14 is Hurwitz. The real parts of both for $s = j\omega$ are non-negative since h was determined such that there were double zeros in the real part of Z_1 for $s = j\omega$ at ± m which forces Z_2 to satisfy the same condition given that the sum of them satisfies it: Also, the real part of Z_2 for $s = j\omega$ has a double zero at the origin, all the zeros in the real parts for $s = j\omega$ are double, and the functions do not change sign in the infinite interval being positive at $\omega = \infty$. Having shown that both functions are P. R. we now show how to reduce them. To realize Z_1 remove the pair of zeros in the $j\omega$ axis by completely removing the corresponding poles of $1/Z_1$. The remainder is realized by inspection. To realize Z_2, remove the zero at the origin by completely removing the corresponding pole from its inverse. The remaining admittance will have zeros on the $j\omega$ axis at $\pm m$. Once these zeros are removed by completely removing the corresponding pole from its inverse, the remainder is realized by inspection. The fact that after the removal of the zero at the origin of Z_2, the remaining function has critical frequencies on the $j\omega$ axis which is not obvious from Equation 17, but can be explained in the following manner:

Let us denote the minimum reactive part of Z_1 by $F_1(s)$ and that of Z_2 by $F_2(s)$. Because of the way we chose the real parts of Z_1 and Z_2 for $s = j\omega$, we must have $F_1(j\omega) = KF_2(j\omega)$, where K is a positive constant. Since Z_1 reduces to a resistance after the removal of all the critical frequencies on the $j\omega$ axis, so must Z_2; thus, it is not by accident that the appropriate coefficients cancel to give critical frequencies on the $j\omega$ axis. The form of the circuit that realizes Equation 14 is shown in Figure 18.

Example 3: The following example illustrates the simplicity with which the method is applied. Consider[5]

$$Z(s) = \frac{s^2 + s + 8}{s^2 + 2s + 2} \tag{21}$$

The real part of $Z(j\omega)$ has double zeros at $\omega = \pm 2$. Using Equation 19 we find

$$h = 2\,\frac{8 + 2 - 2}{2 - 8 + 2} = -4$$

Since $h < 0$ we realize $Z' = 1/Z = \dfrac{s^2 + 2s + 2}{s^2 + s + 8}$; with the new values we find

$$h = 1\,\frac{2 + 8 - 2}{8 - 2 + 2} = 1$$

The impedance leveled and releveled by s + 1 is:

$$Z'(s) = \frac{s^3 + 3s^2 + 4s + 2}{s^3 + 2s^2 + 9s + 8} = \frac{(s^2 + 4)/2}{s^3 + 2s^2 + 9s + 8} + \frac{s^3 + 5s^2/2 + 4s}{s^3 + 2s^2 + 9s + 8},$$

whose realization is shown in Figure 19. The dual of Figure 19 which realizes the original Z(s) of Equation 21 appears in Figure 20.

CONCLUSION

In this paper, the process of impedance leveling was generalized to include, besides constants, ratios of polynomials of s. Through the exhibition of several examples sufficient evidence was given that the generalized process can be a useful analytical tool when applied to the solution of problems in network synthesis. The paper does not pretend to be exhaustive, or to solve new problems. It does not go into synthesis theory more than necessary to illustrate the possibility of attacking problems with this new concept. The objective of the paper was to introduce a tool and then to justify its introduction by showing some of the jobs the tool can do. The impedance leveling process when used in conjunction with other transformations[4] can be an even more powerful tool in circuit theory.

REFERENCES

1. S. Seshu and N. Balabanian, "Transformations of Positive Real Functions," IRE Transactions on Circuit Theory, Theorem 8, (December, 1957).
2. E. A. Guillemin. Synthesis of Passive Networks. John Wiley and Sons, 1957. Pg. 114.
3. E. A. Guillemin. Op. Cit. Pg. 331.
4. M. A. Murray Lasso, "Matrix Transformations in Circuit Theory," S. M. Thesis, Electrical Engineering Department, Massachusetts Institute of Technology, 1962.
5. E. A. Guillemin. Op. Cit. Pg. 396.

APPENDIX

REALIZATION OF THE GENERAL RLC DRIVING POINT IMPEDANCE THROUGH GENERALIZED LEVELING

With a similar argument to the one used for a biquadratic minimum impedance, but adding some complications that arise, we can show that generalized impedance leveling can be used to realize the general RLC driving point impedance.

Given a P. R. function which has been previously minimized and has a double zero in its real part for $s = j\omega$ at $\pm m$

$$Z(s) = \frac{P(s)}{Q(s)} = \frac{m_1 + n_1}{m_2 + n_2}, \tag{A. 1}$$

where m_1 and m_2 are the even parts of the numerator and denominator and n_1 and n_2 the odd parts. Putting the double zero at $\pm m$ in evidence, we write,

$$\mathrm{Ev}\, Z(s) = \frac{m_1 m_2 - n_1 n_2}{m_2^2 - n_2^2} = \frac{(s^2 + m^2)^2 A(-s^2)}{m_2^2 - n_2^2} \tag{A. 2}$$

where

$$A(-s^2) = \frac{m_1 m_2 - n_1 n_2}{(s^2 + m^2)^2} \tag{A. 3}$$

Because the real part for $s = j\omega$ vanishes at a point, we try to extract reactive impedances that will not affect the real part. To accomplish this, level and relevel, Equation A. 1, by $s + h$

$$Z(s) = \frac{(m_1 + n_1)(s + h)}{(m_2 + n_2)(s + h)} = \frac{(s^2 + m^2)\, B(s)}{(m_2 + n_2)(s + h)} + \frac{sC(s)}{(m_2 + n_2)(s + h)} = Z_1 + Z_2 \tag{A. 4}$$

where we have split $Z(s)$ into two parts, one with zeros in the $j\omega$ axis at $\pm m$ and a remainder with a zero at the origin. Let $B(s)$ and h be such that they satisfy

$$\mathrm{Ev}\, Z_1(s) = \frac{(s^2 + m^2)^2 h^2 A(-s^2)}{(m_2^2 - n_2^2)(h^2 - s^2)} = \mathrm{Ev}\ \frac{(s^2 + m^2)\, B(s)(m_2 - n_2)(-s + h)}{(m_2^2 - n_2^2)(h^2 - s^2)} \tag{A. 5}$$

so that the even part of Z_1 has zeros in the same places that the even part of $Z(s)$ has them. Thus, Z_2 must satisfy

$$\mathrm{Ev}\, Z_2(s) = \frac{(s^2 + m^2)^2 A(-s^2)(-s^2)}{(m_2^2 - n_2^2)(h^2 - s^2)} \tag{A. 6}$$

so that $\mathrm{Ev}\, Z = \mathrm{Ev}\, Z_1 + \mathrm{Ev}\, Z_2$. The factor $(s^2 + m^2)/(m_2^2 - n_2^2)(h^2 - s^2)$ is even so we can take it out of the Ev sign in Equation A. 5 and reduce it to

$$\mathrm{Ev}\left[-B(s)(m_2 - n_2)(s - h)\right] = h^2 (s^2 + m^2)\, A(-s^2) \tag{A. 7}$$

Let us call $-B(s) \cdot (s - h)/h^2 = D(s)$ and $(s^2 + m^2)\, A(-s^2) = A'(-s^2)$. The solution of

$$\mathrm{Ev}\,[D(s)(m_2 - n_2)] = A'(-s^2) \tag{A. 8}$$

for D(s) can be done by equating powers of s^2 on both sides. Suppose both P(s) and Q(s) in Equation A. 1 are of degree r, where r is even. (Parallel arguments can be made for other cases.); then, $A(-s^2)$ in Equation A. 3 will be of degree 2r - 4, $A'(-s^2)$ will be of degree 2r - 2. This means that in Equation A. 8 D(s) will, in general, be of degree r - 1. We can write symbolically

$$D(s) = d_{r-1}\, s^{r-1} + d_{r-2}\, s^{r-2} + \ldots + d_o,$$

$$Q(s) = q_r\, s^r + q_{r-1}\, s^{r-1} + \ldots + q_o,$$

$$A'(-s^2) = a'_{2r-2}\, s^{2r-2} + a'_{2r-4}\, s^{2r-4} + \ldots + a'_o.$$

With this nomenclature Equation A. 8 reads:

$$\begin{aligned}(d_{r-2}\, s^{r-2} + d_{r-4}\, s^{r-4} + \ldots + d_o)(q_r\, s^r + q_{r-2}\, s^{r-2} + \ldots + q_o) \\ - (d_{r-1}\, s^{r-1} + d_{r-3}\, s^{r-3} + \ldots + d_1\, s)(q_{r-1}\, s^{r-1} + \ldots + q_1 s) \\ = a'_{2r-2}\, s^{2r-2} + a'_{2r-4}\, s^{2r-4} + \ldots + a'_o,\end{aligned}$$

which gives the set of equations

$$\begin{aligned}-d_{r-1}\, q_{r-1} + d_{r-2}\, q_r &= a'_{2r-2} \\ -d_{r-1}\, q_{r-3} + d_{r-2}\, q_{r-2} - d_{r-3}\, q_{r-1} + d_{r-4}\, q_r &= a'_{2r-4} \\ - & - - - - \\ d_2\, q_o - d_1\, q_1 + d_o\, q_2 &= a'_2 \\ d_o\, q_o &= a'_o\end{aligned} \tag{A. 9}$$

upon equating coefficients of equal powers. In Equation A. 9 the quantities q_i and a'_i are known and we can solve for the d_i to determine D(s). To obtain h and B(s) we determine the roots of D(s) = 0. We show below that D(s) has only one real positive root, this determines h. We then obtain $B(s) = -h^2 D(s)/(s-h)$ with which we can determine Z_1. We determine Z_2 by subtracting Z_1 from Z.

We now prove that both Z_1 and Z_2 are P. R. The real part of $Z_1(j\omega)$ is

$$\text{Re } Z_1(j\omega) = \frac{(m^2 - \omega^2)^2 \, h^2 \, A(\omega^2)}{(m_2^2 - n_2^2)(h^2 + \omega^2)} \qquad \text{(A. 10)}$$

which is non-negative for all ω because $A(\omega^2)$ comes from Equation A. 2 and we assumed that $Z(s)$ of Equation A. 1 was P. R. Z_2 has real part for $s = j\omega$

$$\text{Re } Z_2(j\omega) = \frac{(m^2 - \omega^2)^2 \, A(\omega^2)\, \omega^2}{(m_2^2 - n_2^2)(h^2 + \omega^2)} \qquad \text{(A. 11)}$$

which is also non-negative for all ω. To prove that Z_1 and Z_2 are Hurwitz, we need to prove that a real positive h can always be found. Let us rewrite Equation A. 7 in the form:

$$\text{Re}\left[-B(j\omega)(j\omega - h)\, Q(-j\omega)\right] = - \left|Q(j\omega)\right|^2 \text{Re}\left[B(j\omega)(j\omega - h)/Q(j\omega)\right]$$

$$= (-\omega^2 + m^2)\, A(\omega^2)\, h^2 \qquad \text{(A. 12)}$$

The right side of Equation A. 12 changes sign once in the interval $0<\omega<\infty$ and has negative leading coefficient , since $A(\omega^2)$ is non-negative for all ω and has zeros other than $\pm m$. Accordingly, the shape of $\text{Re}\left[B(j\omega)(j\omega - h)/Q(j\omega)\right]$ is similar to the one shown in Figure A. 1. $B(s) \cdot (s - h)$ can have at most one zero in the right half plane because there is only one change of sign in $0 < \omega < \infty$. Thus, the shape of the curve can be due to either of two things:

1. $B(s) \cdot (s - h)$ has $h > 0$ and $B(s)$ Hurwitz, or
2. $B(s) \cdot (s - h)$ is all Hurwitz but $B(s)$ has a negative multiplier associated with it.

Case 1. solves the problem since then both Z_1 and Z_2 will be P. R. To test if we are in case 1., we have to insure that the sign of the leading coefficient of $B(s)$ is positive. Dividing Equation A. 12 by $h^2 \left|Q(j\omega)\right|^2$ and comparing with Equation A. 3, we obtain

$$- \frac{1}{h^2} \text{Re}\left[B(j\omega)(j\omega - h)/Q(j\omega)\right] = (-\omega^2 + m^2) A(\omega^2)/\left|Q(j\omega)\right|^2 =$$

$$= \text{Re}\left[Z(j\omega)/(m^2 - \omega^2)\right], \qquad \text{(A. 13)}$$

Thus,

$$\text{Re } Z(j\omega) = - \text{Re}\left[(m^2 - \omega^2) B(j\omega)(j\omega - h)/Q(j\omega)\right](1/h^2) \qquad \text{(A. 14)}$$

If we now perform the division $(s^2 + m^2)B(s)(s - h)/Q(s)h^2$ recalling that $B(s) \cdot (s - h)$ is of degree $r - 1$, while $Q(s)$ is of degree r, we obtain

$$\frac{(s^2 + m^2)B(s)(s - h)}{h^2\, Q(s)} = ks + \frac{R(s)}{Q(s)} \qquad (A.\ 15)$$

The sign of k will be the sign of the leading coefficient of B(s), since Q(s) is Hurwitz. Taking real parts for $s = j\omega$ on both sides of Equation A. 15 and comparing with Equation A. 14, we find that $R(s) = -P(s)$ using the fact that an impedance is completely characterized by its real part when it is minimum reactive. Thus, from Equation A. 15 we obtain

$$\frac{1}{h^2} B(s)(s - h) = \frac{Q(s)[ks - P(s)/Q(s)]}{s^2 + m^2} = [ksQ(s) - P(s)]/(s^2 + m^2) \qquad (A.\ 16)$$

therefore, the numerator must contain $s^2 + m^2$ as a factor, which implies $ksQ(s) = P(s)$ for $s = jm$, or $k = P(jm)/jmQ(jm) = Z(jm)/jm$. Therefore, if $Z(jm)/jm$ is positive, the leading coefficient of B(s) will be positive and so will h. If, however, $Z(jm)/jm$ is negative, we realize $Z'(s) = 1/Z(s)$ and we will get a positive h because $Z'(jm)/jm$ will be positive. The network dual to the realization of $Z'(s)$ will realize Z(s). Thus, we have proved that it is possible in all cases with the proper choice to obtain a positive h and thus both Z_1 and Z_2 are P. R.

Finally, to realize Z_1, remove the poles at ±m and at ∞ from $1/Z_1$ obtaining

$$Z_1(s) = \frac{1}{\dfrac{as}{s^2 + m^2} + bs + \dfrac{1}{Z_3(s)}} \qquad (A.\ 17)$$

Likewise, $Z_2(s)$ has the following expansion

$$Z_2(s) = \frac{1}{\dfrac{c}{s} + \dfrac{1}{\dfrac{ds}{s^2 + m^2} + Z_4(s)}} \qquad (A.\ 18)$$

The appearance of the $s^2 + m^2$ in $Z_2(s)$, in spite of the fact that it was not apparent in Equation A. 4, can be explained. The fact that since both $Z_1(s)$ and $Z_2(s)$ have the same real part for $s = j\omega$ except for a factor ω^2/h^2 (see Equations A. 10 and A. 11), once the critical frequencies are removed from the $j\omega$ axis the remainders must differ by a multiplying constant, that is, $Z_3(s) = KZ_4(s)$ in Equations A. 17 and

A. 18. This is so because two minimum reactive impedances are completely specified by their real parts. Since both Z_1 and Z_2 must have double zeros in their real parts for $s = jm$, these zeros manifest themselves as shown in Equations A. 17 and A. 18. The remainders Z_3 and Z_4 which are of degree $r - 2$ can be treated in the same way, and so on, until a complete realization is achieved. The circuits obtained are the same circuits as the ones obtained using the Bott-Duffin method.*

ACKNOWLEDGEMENT

The research reported in this paper was supported by the National Science Foundation Grant GP376, M.I.T. DSR 9323.

---

* E. A. Guillemin obtains this result augmenting the real part of the impedance in "New Methods of Driving Point and Transfer Impedance Synthesis". Procedings of the Symposium on Modern Network Synthesis. Polytechnic Institute of Brooklyn, 1956.

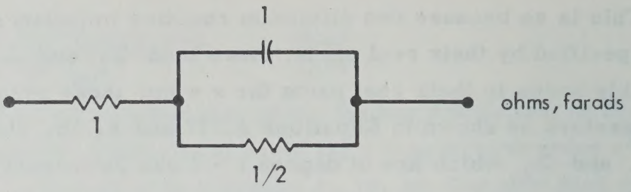

1
1
1/2
ohms, farads

Fig. 1

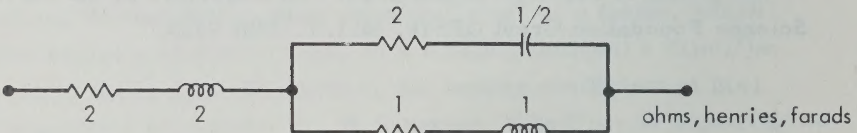

2
1/2
2
2
1
1
ohms, henries, farads

Fig. 2

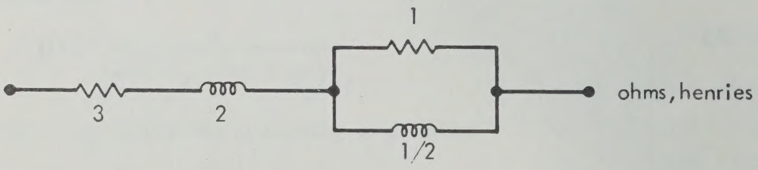

1
3
2
1/2
ohms, henries

Fig. 3

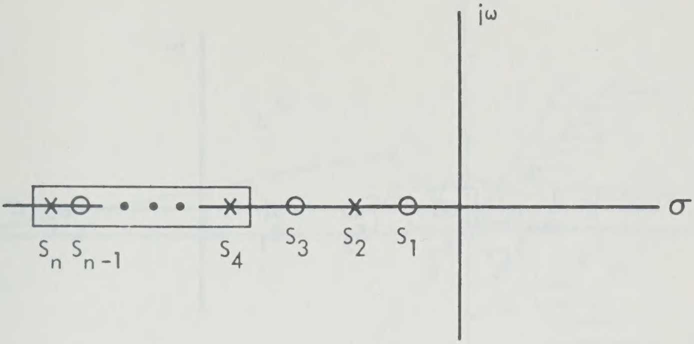

jω
σ
S_n
S_{n-1}
S_4
S_3
S_2
S_1

Fig. 4

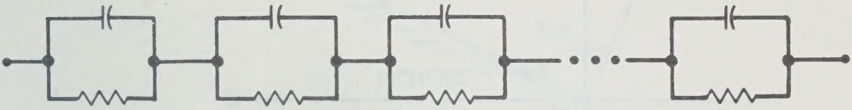


Fig. 5

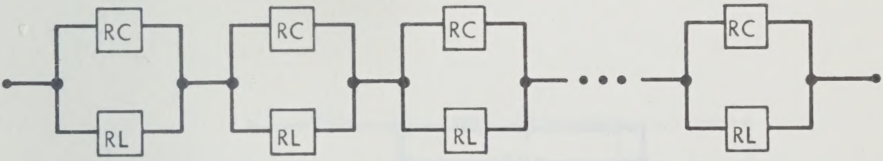

RC
RC
RC
RC
RL
RL
RL
RL

Fig. 6

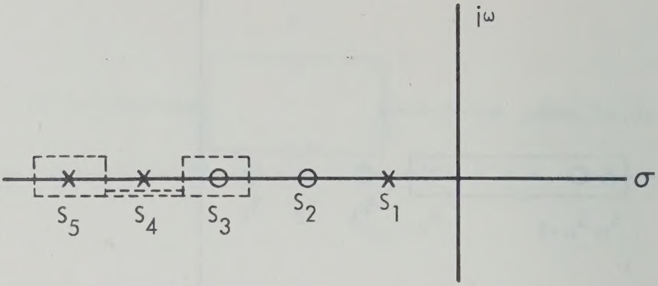

jω
σ
S_5
S_4
S_3
S_2
S_1

Fig. 7

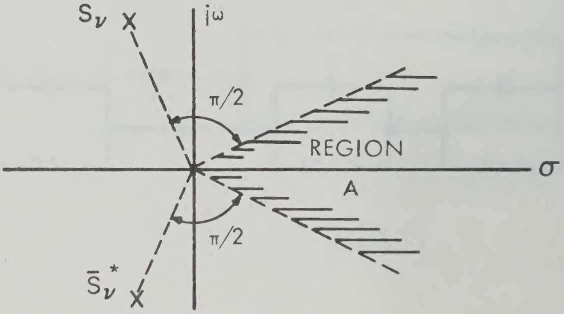

S_ν
jω
π/2
REGION
A
σ
π/2
$\bar{S}_\nu^*$

Fig. 8

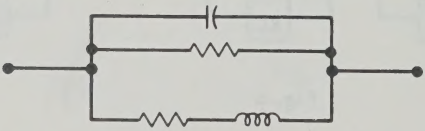


Fig.9

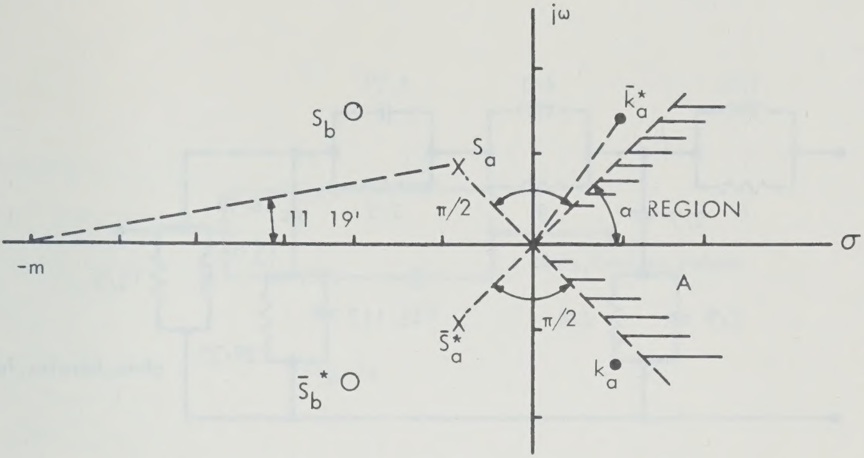
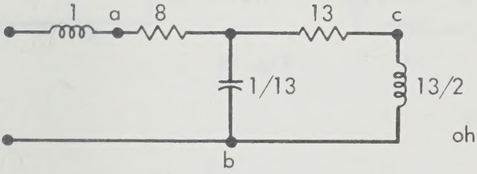
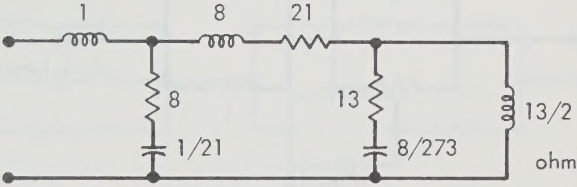

jω
S_b
S_a
$\bar{k}_a^*$
11
19'
π/2
α
REGION
A
σ
-m
π/2
$\bar{S}_a^*$
$\bar{S}_b^*$
k_a

Fig. 10


1
a
8
13
c
1/13
13/2
b

ohms, henries, farads

Fig. 11


1
8
21
8
13
13/2
1/21
8/273

ohms, henries, farads

Fig. 12

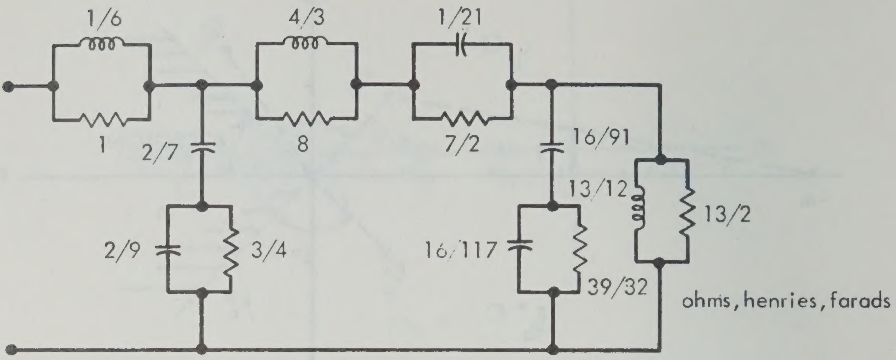

1/6
4/3
1/21
1
2/7
8
7/2
16/91
13/12
13/2
2/9
3/4
16/117
39/32
ohms, henries, farads

Fig. 13

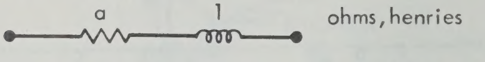

a
1
ohms, henries

Fig. 14

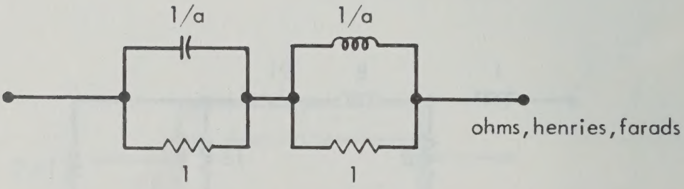

1/a
1/a
ohms, henries, farads
1
1

Fig. 15

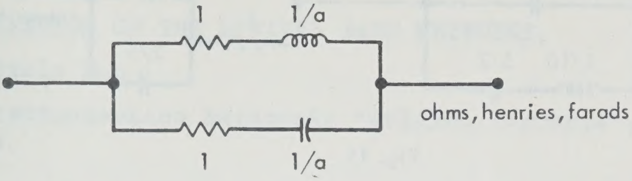
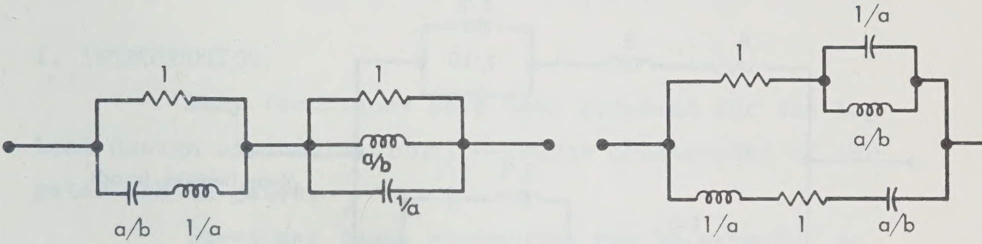

1
1/a
ohms, henries, farads
1
1/a

Fig. 16


1
1
a/b
1/a
a/b
1/a
1/a
1
a/b
1/a
1
a/b

(a) (b)

Fig. 17

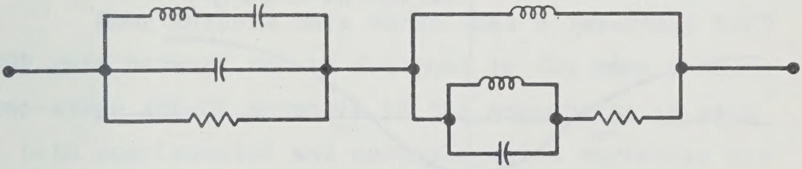


Fig. 18

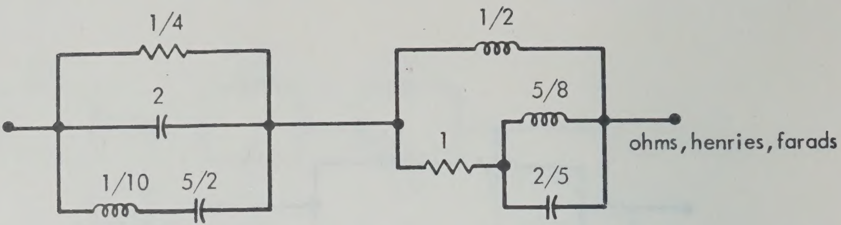

1/4
2
1/10
5/2
1/2
5/8
1
2/5
ohms, henries, farads

Fig. 19

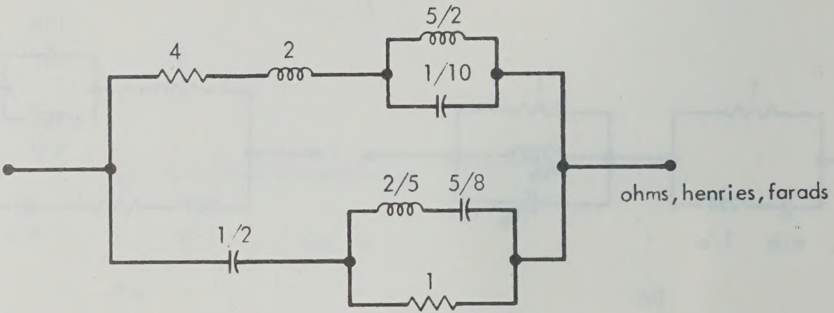

4
2
5/2
1/10
2/5
5/8
1/2
1
ohms, henries, farads

Fig. 20

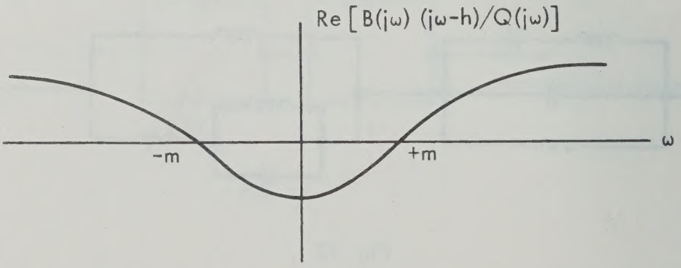


Fig. A.1

THE DETERMINATION OF THE MINIMAL TANT NETWORKS.

Angelo Raffaele Meo
Istituto Elettrotecnico Nazionale "Galileo Ferraris", Torino, Italy.

I. INTRODUCTION.

Many techniques have been proposed for the logical design of combinational circuits constructed of AND gates and OR gates.

Sometimes these procedures can be extended to apply to networks of NAND or NOR gates(1), but the use of these gates often gives rise to problems for which new techniques are required. The problem considered in this paper is one of those in which it is necessary to make reference to specific properties of NAND or NOR gates for the logical design of the minimal network.

Some authors have shown that a two-stage NAND or NOR gate network can be designed in the same fashion as two-stage AND-OR networks if the assumption is made that both complemented and uncomplemented variables are available(2)(3). If only the uncomplemented variables are present, the assigned function cannot as a rule be implemented by a two-stage network; however every function can always be realized by using a three-stage network. A net work, of this type, that is a three-stage AND-NOT network with true or uncomplemented inputs, has been defined as a TANT network.

An interesting procedure for designing TANT networks has been described by E.J. McCluskey(4); however it is not guaranteed to result in networks having a minimum number of gates. In this paper some properties of minimal TANT networks will be discussed, and a simple algebraic procedure for determining them will be described.

II. BASIC DEFINITIONS.

A. TANT NETWORKS.

In this paper attention will be restricted to single-output networks of NAND gates which contain three levels or stages of gates between input and output, and which have only uncomplemented variables as inputs. The restriction to three stages corresponds to a requirement of maximum speed in forming the output, while the restriction to NAND networks is not substantial, since the rules for extending a set of results or properties concerning NAND networks to NOR networks are well known. Be it remembered that a NAND or AND-NOT gate is an elementary combinational circuit whose output is 0 only when all the inputs are 1, and is 1 elsewhere, while a NOR or OR-NOT gate is a circuit whose output is 1 only when all the inputs are 0, and is 0 elsewhere. Therefore a NOR network implementing an assigned function T corresponds to the NAND network implementing the complementary function $\overline{T}$, in as much as the former can be obtained from the latter by substituting the NAND gates with NOR gates and by complementing the input variables. In particular, this rule is valid for the minimal TANT networks of the two dual kinds.

The general structure of a TANT network is illustrated by Figure 1. The gate from which the output is derived will be called the "first-stage (or output) gate", the gates whose outputs are connected to the inputs of the first-stage gate will be called "second-stage gates" and the remaining gates will be called "third-stage gates".

B. TANT EXPRESSIONS.

A TANT network can be described with a concise expression by using the symbol of the NAND operator. For example, the network of Figure 1 corresponds to the following expression:

$$\left\{\left[(/a)/(b/c)\right] / \left[a/(b/c)/(d/e/f)\right] /g\right\} \sim T \qquad (1)$$

In this expression the variables corresponding to the inputs of the first-stage gate are called "first-stage variables", and analogously the variables corresponding to the inputs of the second- and third-stage gates are called "second- and third-stage variables" respectively.

It is easy to transform an expression containing the NAND operator in a conventional AND-OR expression. In particular a third-order NAND expression can be transformed into the corresponding "sum of products of sums of literals"

i) by substituting the first- and third-stage operators with sum operators and the second-stage operators with multiplication operators;

ii) by complementing the first- and third-stage variables and keeping the remaining ones unchanged.

For example, expression (1) is equivalent to the following expression of T:

$$\left[\bar{a}\cdot(\bar{b}+\bar{c})+a\cdot(\bar{b}+\bar{c})\cdot(\bar{d}+\bar{e}+\bar{f})+\bar{g}\right]\sim T \qquad (2)$$

A"sum of products of sums of literals" corresponding to a TANT network implementing a function T will be called "TANT expression of T". In particular the TANT expression corresponding to the TANT network implementing T and having the minimum number of gates will be called "minimal TANT expression of T".

Be it noted that more than one network correspond to a TANT expression, since the output of a third--stage gate may be applied to more than one gate of the second- or even first-stage. For example, both the networks of Figures 1 and 2 correspond to the same expression. However, since the minimal TANT network is sought for, assumption will be made in this paper that the network corresponding to a given expression is that which has the minimum number of gates.

III. SOME PROPERTIES OF TANT NETWORKS.

In this Section some interesting properties of TANT networks will be proved. For the sake of simplicity, they will be discussed in terms of TANT expressions.

THEOREM 1.

All the third-stage variables of every TANT expression must be complemented.

Proof.

The input variables of a TANT network are uncomplemented; therefore, by rule ii) of Section II, the third-stage variables of the corresponding TANT expression will be complemented.

Notice that the first- and second-stage variables may be either complemented or uncomplemented; however while the complemented first-stage and the uncomplemented second-stage variables correspond to direct connections between the inputs and the first- and second-stage gates, the complementation of the uncomplemented first-stage variables and of the complemented second-stage variables requires the use of one-input gates (of the type of the third-stage gate complementing input a in Figure 1).

Definitions.

1.- Each addendum of a TANT expression t^* will be called "second-order component expression of t". For example, the second-order components expressions of expression (2) are the following:

$$ps_1F_1 = \left[\bar{a}\cdot(\bar{b}+\bar{c})\right]$$

$$ps_2F_2 = \left[a\cdot(\bar{b}+\bar{c})\cdot(\bar{d}+\bar{e}+\bar{f})\right]$$

$$ps_3F_3 = \left[\bar{g}\right]$$

2.- Functions F_1, F_2, ..., F_j, ... expressed by the second-order component expressions of t will be called "component functions of t" (CFt). The component functions of expression (2) are the following:

---

* Capital and small letters will be used for functions and expressions respectively; the symbol = will denote the identity of expressions, the symbol $\sim$ the logical equivalence of expressions or functions. sF and pF will be used to respectively denote expressions of the types "sum of literals" or "product of literals", logically equivalent to F; analogously, the symbols spF and psF will be used for "sum of products of literals" and "product of sums of literals" expressions respectively. Also, third-order expressions will be denoted with the symbols

./.

$$F_1 \sim \left[\bar{a}\cdot(\bar{b}+\bar{c})\right] \sim \left[\bar{a}\bar{b}+\bar{a}\bar{c}\right]$$

$$F_2 \sim \left[a\cdot(\bar{b}+\bar{c})\cdot(\bar{d}+\bar{e}+\bar{f})\right] \sim \left[a\bar{b}\bar{d}+a\bar{b}\bar{e}+a\bar{b}\bar{f}+a\bar{c}\bar{d}+a\bar{c}\bar{e}+a\bar{c}\bar{f}\right]$$

$$F_3 \sim \left[\bar{g}\right]$$

Since many of the properties of second-order expressions which will be used afterwards are more usually known in terms of sp expressions than of ps expressions, it is necessary to introduce the complementary definitions of the ones given above.

1'.- By "complement of a second-order component expression ps_jF_j of t" will be meant the expression $sp_j\bar{F}_j$ which can be obtained from ps_jF_j by substituting the OR operators with AND operators, and viceversa, and by complementing all the variables. So the complements of the second-order component expressions of expression (2) are the following:

$$sp_1\bar{F}_1 = \left[a + bc\right]$$

$$sp_2\bar{F}_2 = \left[\bar{a} + bc + def\right]$$

$$sp_3\bar{F}_3 = \left[g\right]$$

2'.- Functions $\bar{F}_1$, $\bar{F}_2$, ..., $\bar{F}_j$, ... expressed by the complements of second-order component expressions will be called "complements of the component functions" (CCF).

---

spsF and pspF, and in particular $sps_{MT}F$ will be the minimal TANT expression of function F.

THEOREM 2

All the variables of every factor containing more than one literal of every second-order component expression must be complemented.

THEOREM 2'

The complement of every second-order component expression is of the following type:

$$sp_j\bar{F}_j = (\bar{x}_{a_1} + \bar{x}_{a_2} + \dots + \bar{x}_{a_l} + x_{b_1} + x_{b_2} + \dots + x_{b_m} + x_{c_{11}} x_{c_{12}} \dots x_{c_{1r}} + x_{c_{21}} x_{c_{22}} \dots x_{c_{2s}} + \dots + x_{c_{t1}} x_{c_{t2}} \dots x_{c_{tv}}) \qquad (3)$$

where all the variables of every addendum containing more than one literal* must be uncomplemented.

Theorem 2 and 2' follow from Theorem 1 directly.

THEOREM 3.

If a prime implicant of the complement of a component function has more than one literal, all its variables are uncomplemented.

Proof.

If $\bar{F}_j$ is a CCF, it has an expression of type (3). By Quine's Iterated Consensus Theorem$^{(5)}$ all the prime implicants of $\bar{F}_j$ can be obtained with a repeated application of the consensus operation** to the addenda of expression (3).

---

* Literal is a variable either complemented or uncomplemented. It will be denoted with the symbol x^* .

** The consensus of two terms A and B is that term C which exists only if A and B contain one and only one variable x_0 being complemented in the one and uncomplemented in the other, and which contains all the literals of A and B, common and not common, minus

The consensus of a term $\bar{x}_{a_i}$ with a term x_{b_j} exists if $x_{a_i} = x_{b_j}$; however, in this case, $\bar{\bar{F}}_j \sim 1$ and 1 is the only prime implicant of $\bar{\bar{F}}_j$. The consensus of a term x_{b_j} with a term $x_{c_{k1}}\ x_{c_{k2}} \ldots x_{c_{kh}}$ cannot exist since all the variables contained in them are uncomplemented. For the same reason, the consensus of a term $\bar{x}_{a_i}$ with a $\bar{x}_{a_j}$, or of a x_{b_i} with a x_{b_j}, or of a $x_{c_{i1}}\ x_{c_{i2}} \ldots x_{c_{ih}}$ with a $x_{c_{j1}}$ $x_{c_{j2}} \ldots x_{c_{jk}}$ cannot exist. A consensus may only be given by a pair made of a term $A = \bar{x}_{a_i}$ and a term $C = x_{c_{j1}} x_{c_{j2}} \ldots x_{c_{jk}}$; in this case, one of the variables of C, e.g. $x_{c_{j1}}$, must be $= x_{a_i}$, and the consensus will be the term $x_{c_{j2}} \ldots x_{c_{jk}}$ which contains all the variables of C minus x_{a_i}. Therefore the repeated application of the consensus operation to the addends of expression (3) will always give terms containing only uncomplemented variables. It follows that either a prime implicant of $\bar{\bar{F}}_j$ contains only one literal or all its variables are uncomplemented.

Definitions.

An addendum A of the complement of a second-order component expression $sp_j\bar{\bar{F}}_j = (A+B+C+\ldots+Z)$ will be called a "superfluous addendum" if it implies the sum $(B+C+\ldots+Z)$ of the remaining addenda. Obviously a superfluous addendum can always be stricken off, since its removal does not change the logical content of the considered expression and does not increase the number of gates of the corresponding network.

---

x_o. For example, the consensus of $A = ab\bar{c}$ and $B = bc\bar{d}\bar{e}$ is $C = ab\bar{d}\bar{e}$.

The complement of a second-order expression not containing superfluous addenda will be called "irredundant". Also the second-order component expressions, which are the complements of irredundant expressions, will be called "irredundant".

THEOREM 4.

Consider the complement of a second-order component expression

$$sp_j\bar{F}_j = (\bar{x}_{a_1} + \bar{x}_{a_2} + \dots + \bar{x}_{a_e} + A_1 + A_2 + \dots + A_m) \quad (4)$$

where $A_1, A_2, \dots, A_m$ denote products of uncomplemented variables. For every prime implicant P of $\bar{F}_j$ there is a corresponding addendum in sp_jF_j, which either coincides with P or is the product of P by one or more than one of variables $x_{a_1}, x_{a_2}, \dots, x_{a_e}$.

Proof.

By Quine's Iterated Consensus Theorem every prime implicant of $\bar{F}_j$ can be obtained with a repeated application of the consensus operation to the addenda of expression (4). We have proved that a consensus may only be given by a pair made of a term $\bar{x}_{a_i}$ and a term A_h, when one of the variables of A_h is coincident with x_{a_i}, in which case the consensus will be that term which contains all the variables of A_h minus x_{a_i}. Therefore for every prime implicant P of $\bar{F}_j$, there must be a corresponding addendum in $sp_j\bar{F}_j$ which either coincides with P or contains all the variables of P plus some of literals x_{a_1}, $x_{a_2}, \dots, x_{a_e}$.

THEOREM 5.

In an irredundant complement of a second-order component expression $sp_j\bar{F}_j$ *every addendum is either a prime implicant of* $\bar{F}_j$ *or the product of a prime implicant by one or more than one of the variables* $x_{a_1}, x_{a_2}, \dots, x_{a_e}$ *which are complemented in the expression.*

Proof.

Suppose that in an irredundant complement of a second-order component expression:

$$sp_j\bar{F}_j = (\bar{x}_{a_1} + \bar{x}_{a_2} + \dots + \bar{x}_{a_e} + A_1 + A_2 + \dots + A_m) \qquad (4)$$

addendum A_i is neither a prime implicant of $\bar{F}_j$ nor the product of a prime implicant of $\bar{F}_j$ by some of variables $x_{a_1}, x_{a_2}, \dots, x_{a_e}$. Since A_i is not a prime implicant, there must be a prime implicant A_j^* which is implied by A_i. Also, by Theorem 4, there must be a corresponding addendum A_j in $sp_j\bar{F}_j$ which either coincides with A_j^* or is the product of A_j^* by one or more than one of variables $x_{a_1}, x_{a_2}, \dots, x_{a_e}$. Therefore A_i implies $(A_j + \bar{x}_{a_1} + \bar{x}_{a_2} + \dots + \bar{x}_{a_e})$. Hence A_i is superfluous.

THEOREM 6.

If $\bar{F}_j$ *is a CCF of a TANT expression of a function T, then* $\bar{T}$ *implies* $\bar{F}_j$.

Proof.

Since T is the sum of the second-order component expressions, each of these implies T. Theorem 6 expresses this peoperty in complemented terms.

Definitions.

Let $T(x_1, x_2, \dots, x_n)$ be an assigned function.

Consider the following functions:

$$G_{1,0,0,\ldots,0,0} \sim (\bar{x}_1 + \bar{T})$$

$$G_{0,1,0,\ldots,0,0} \sim (\bar{x}_2 + \bar{T})$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$G_{0,0,0,\ldots,0,1} \sim (\bar{x}_n + \bar{T})$$

$$G_{1,1,0,\ldots,0,0} \sim (\bar{x}_1 + \bar{x}_2 + \bar{T})$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$G_{0,0,0,\ldots,1,1} \sim (\bar{x}_{n-1} + \bar{x}_n + \bar{T})$$

$$G_{1,1,1,\ldots,0,0} \sim (\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{T})$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$G_{1,1,1,\ldots,1,1} \sim (\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \ldots + \bar{x}_{n-1} + \bar{x}_n + \bar{T})$$

which can be obtained by adding $\bar{T}$ to the sum of the elements of every possible subset of complemented variables.

Each of these functions can be easily expressed as a sum of all its prime implicants:

$$G_{\alpha,\beta,\gamma,\ldots,\omega,\xi} \sim (\bar{x}_{a_1} + \bar{x}_{a_2} + \ldots + \bar{x}_{a_r} + B_1C_1 + B_2C_2 + \ldots + B_sC_s + D_1 + D_2 + \ldots + D_t) \qquad (5)$$

where the C_i's and the D_j's denote products of uncomplemented variables, and the B_k's denote products of complemented variables. The definition "fundamental function (FF) of T" will be used for the complement of every function of the following type:

$$(G_{\alpha,\beta,\gamma,\ldots,\omega,\xi})_h \sim \left[\bar{x}_{a_1} + \bar{x}_{a_2} + \ldots + \bar{x}_{a_r} + (C_1)_h + (C_2)_h + \ldots + (C_s)_h + (D_1)_h + (D_2)_h + \ldots + (D_t)_h\right] \qquad (6)$$

where $(C_i)_h$ and $(D_j)_h$ are products of uncomplemented variables including or coincident with C_i and D_j

respectively. Also, every function of type (6) will be called "complement of a fundamental function (CFF) of T".

THEOREM 7.

There is always one minimal TANT expression of T, in which each component function is a fundamental function of T.

Proof.

Since every CCF has an expression of type (3), it is implied by the sum of the elements of a subset Ω of complemented variables $(\bar{x}_{a_1} + \bar{x}_{a_2} + \dots + \bar{x}_{a_e})$; besides, by Theorem 6, every CCF is implied by $\bar{T}$. Therefore every CCF is implied by that $G_{\alpha,\beta,\gamma,\dots\dots,\omega,\xi}$ which is the sum of $\bar{T}$ plus the elements of subset Ω.

Let t be a minimal TANT expression of a function T:

$$t = sps_{MT}T = (ps_1F_1 + ps_2F_2 + \dots + ps_jF_j + \dots + ps_zF_z) \quad (7)$$

and suppose a component function F_j is not a fundamental function of T. Let G^* be that $G_{\alpha,\beta,\gamma,\dots\dots,\omega,\xi}$ which implies $\bar{F}_j$.

Since every prime implicant of G^* must imply or coincide with a prime implicant of $\bar{F}_j$, and since, by Theorem 3, all the variables of every prime implicant of $\bar{F}_j$ containing more than one literal must be uncomplemented, then $\bar{F}_j$ will be of the following type:

$$\bar{F}_j \sim \Big[\bar{x}_{a_1} + \bar{x}_{a_2} + \dots + \bar{x}_{a_e} + (C_1)^* + (C_2)^* + \dots + (C_s)^* + \\ + (D_1)^* + (D_2)^* + \dots + (D_t)^* + (E_1)^* + (E_2)^* + \dots + \\ + (E_v)^*\Big] \quad (8)$$

where $(C_i)^*$ and $(D_j)^*$ are products of uncomplemented variables including or coincident with the terms C_i and D_j of G^* (see expression (5) of G^*), while terms $(E_k)^*$ are other products of uncomplemented variables.

By comparing expression (8) with expression (6) one can conclude that $\bar{F}_j$ is the sum of the complement of a fundamental function F_j' of T plus some products of uncomplemented variables:

$$\bar{F}_j \sim \left[\bar{F}_j' + (E_1)^* + (E_2)^* + \dots + (E_v)^*\right]$$

But the addenda $(E_1)^*, (E_2)^*, \dots, (E_v)^*$ can be eliminated. In fact:

a) if expression (7) implements function T, also expression

$$t = (ps_1F_1 + ps_2F_2 + \dots + ps_jF_j' + \dots ps_zF_z)$$

implements T, since F_j' implies T and is implied by F_j;

b) the cost of the irredundant expressions of $\bar{F}_j$ is superior to the cost of the irredundant expressions of $\bar{F}_j'$, since, by Theorem 4 and 5, the complement of every second-order component expression contains as many addenda as are the prime implicants of the corresponding CCF.

Therefore expression (7) cannot be a minimal TANT expression of T.

IV. THE SYNTHESIS PROCEDURE.

The theorems proved in previous Section III suggest the following procedure for the determination of at least one of the minimal TANT networks implementing as assigned function $T(x_1, x_2, \dots, x_n)$. The procedure consists of the following steps of computation.

STEP 1.

Determination of all the fundamental functions of T.

The definition which has been enunciated in Section III shows an easy direct way to determine all the fundamental functions of T.

STEP 2.

Search for all the irredundant sums of fundamental functions which implement T.

By Theorem 7, there is always one minimal TANT expression of T, in which each component function is a fundamental function of T. Besides, the minimal TANT expressions are irredundant; that is, no component function implies the sum of the remaining ones. Therefore there is always one minimal TANT expression which corresponds to an irredundant sum of fundamental functions.

Well known procedures make it possible to determine all the irredundant sums of fundamental functions which implement T; at least one of the minimal TANT expressions of T corresponds to one of these irredundant sums.

STEP 3.

Minimization of the TANT expression corresponding to every irredundant sum of fundamental functions.

Suppose we know the component functions F_1, F_2, ..., F_j, ..., F_p of an minimal TANT expression. Let $\bar{x}_{a_1}$, $\bar{x}_{a_2}$, ..., $\bar{x}_{a_e}$, A_1, A_2, ..., A_m (where A_1, A_2, ..., A_m denote products of uncomplemented variables) be the prime implicants of $\bar{F}_j$. By Theorems 4 and 5, in the complement of the second-order component expression corresponding to

$\bar{F}_j$ there are as many addenda as are the prime implicants of the corresponding function. But we do not know whether the addendum corresponding to A_i will be A_i itself or $x_{a_1}A_i$ or $x_{a_1}\,x_{a_2}\,A_i$ or the product of A_i by some other subset of variables $x_{a_1}, x_{a_2}, \dots, x_{a_e}$. Therefore for every irredundant sum of fundamental functions, it is necessary to find the minimal corresponding expression. Obviously the minimal TANT expression of the assigned function will be the minimal one among the minimal expressions corresponding to every irredundant sum of fundamental functions. The problem of determining the minimal expression corresponding to an assigned sum of functions offers no difficulty, and can be solved by a number of simple procedures. One of these will be described in the example which is discussed in the next Section.

V. AN EXAMPLE.

Let the assigned function be $T \sim (\bar{a}b\bar{c}+\bar{a}\bar{b}c+\bar{b}c\bar{d}+a\bar{b}\bar{c}+a\bar{b}\bar{d})$.

STEP 1.

Determination of the fundamental functions.

The complementary function of the assigned function is

$$\bar{T} \sim (\bar{a}\bar{b}\bar{c} + bc + ab + acd)$$

Therefore the complements of the fundamental functions:

$$G_{1,0,0,0} \sim (\bar{a} + \bar{T}) \sim (\bar{a} + b + cd)$$

from which we obtain:

$$(G_{1,0,0,0})_1 \sim (\bar{a} + b + c)$$

$$(G_{1,0,0,0})_2 \sim (\bar{a} + b + d);$$

$G_{0,1,0,0} \sim (\bar{b} + \bar{T}) \sim (\bar{b} + c + a);$

$G_{0,0,1,0} \sim (\bar{c} + \bar{T}) \sim (\bar{c} + b + ad)$

from which we obtain:

$(G_{0,0,1,0})_1 \sim (\bar{c} + b + a)$

$(G_{0,0,1,0})_2 \sim (\bar{c} + b + d);$

$(G_{1,1,0,0}) \sim (\bar{a} + \bar{b} + \bar{T}) \sim 1;$

$(G_{1,0,1,0}) \sim (\bar{a} + \bar{c} + \bar{T}) \sim (\bar{a} + \bar{c} + b + d);$

$(G_{0,1,1,0}) \sim (\bar{b} + \bar{c} + \bar{T}) \sim 1.$

Notice that the terms of type $G_{\alpha,\beta,\gamma,1}$ have not been considered. In fact, in a fundamental function of type dA literal d is superfluous if the assigned function T is antitonic in d.

Therefore the fundamental functions of T are the following:

$S_1 \sim [a\bar{b}\,(\bar{c} + \bar{d})]$

$S_2 \sim [a\bar{b}\bar{c}]$

$S_3 \sim [a\bar{b}\bar{d}]$

$S_4 \sim [\bar{a}b\bar{c}]$

$S_5 \sim [\bar{b}c\,(\bar{a} + \bar{d})]$

$S_6 \sim [\bar{a}\bar{b}c]$

$S_7 \sim [\bar{b}c\bar{d}]$

$S_8 \sim [a\bar{b}c\bar{d}]$

STEP 2.

Selections of fundamental functions.

The following irredundant sums of fundamental functions can be easily determined by solving the covering problem with one of the conventional procedures:

$S_1+S_4+S_5$, $S_2+S_4+S_5$, $S_1+S_4+S_6$, $S_2+S_3+S_4+S_6$, $S_2+S_4+S_6+S_7$, $S_2+S_4+S_6+S_8$.

STEP 3.

Minimization of the expressions corresponding to the irredundant sums.

Consider the irredundant sum

$$(S_1+S_4+S_5) \sim \left[a\bar{b}\cdot(\bar{c}+\bar{d}) + \bar{a}b\bar{c} + \bar{b}c\cdot(\bar{a}+\bar{d})\right]$$

Every fundamental function corresponds to more than one possible expressions; for example, function S_1 corresponds to the following expressions:

$$s_{11} = \left[a \cdot \bar{b} \cdot (\bar{c}+\bar{d})\right]$$

$$s_{12} = \left[a \cdot (\bar{a}+\bar{b})\cdot(\bar{c}+\bar{d})\right]$$

$$s_{13} = \left[a \cdot \bar{b} \cdot (\bar{a}+\bar{c}+\bar{d})\right]$$

$$s_{14} = \left[a \cdot (\bar{a}+\bar{b})\cdot(\bar{a}+\bar{c}+\bar{d})\right]$$

and, analogously, S_4 corresponds to:

$$s_{41} = \left[b \cdot \bar{a} \cdot \bar{c}\right]$$

$$s_{42} = \left[b \cdot (\bar{a}+\bar{b})\cdot \bar{c}\right]$$

$$s_{43} = \left[b \cdot \bar{a} \cdot (\bar{b}+\bar{c})\right]$$

$$s_{44} = \left[b \cdot (\bar{a}+\bar{b})\cdot(\bar{b}+\bar{c})\right]$$

and S_5 corresponds to

$$s_{51} = \left[c \cdot \bar{b} \cdot (\bar{a}+\bar{d})\right]$$

$$s_{52} = \left[c \cdot (\bar{b}+\bar{c})\cdot(\bar{a}+\bar{d})\right]$$

$$s_{53} = \left[c \cdot \bar{b} \cdot (\bar{a}+\bar{c}+\bar{d})\right]$$

$$s_{54} = \left[c \cdot (\bar{b}+\bar{c})\cdot(\bar{a}+\bar{c}+\bar{d})\right]$$

Every possible sum of an expression s_{1i} plus an s_{4j} plus an s_{5k} corresponds to the irredundant sum $(S_1 + S_4 + S_5)$; however every different $(s_{1i} + s_{4j} + s_{5k})$ requires a different number of gates for its implementation. For example, expression

$$s_{11}+s_{41}+s_{51} = \left[a\bar{b}\,(\bar{c}+\bar{d})+b\bar{a}\bar{c}+c\bar{b}\,(\bar{a}+\bar{d})\right]$$

requires 9 gates (see figure 3), while expression

$$s_{14}+s_{44}+s_{54} = \left[a(\bar{a}+\bar{b})(\bar{a}+\bar{c}+\bar{d})+b(\bar{a}+\bar{b})(\bar{b}+\bar{c})+c(\bar{b}+\bar{c})(\bar{a}+\bar{c}+\bar{d})\right] \qquad (9)$$

requires only 7 gates (see Figure 4), since expression $(\bar{a}+\bar{b})$, being common to s_{14} and s_{44}, is implemented only once, as also $(\bar{b}+\bar{c})$, which is common to s_{44} and s_{54}, and $(\bar{a}+\bar{c}+\bar{d})$, which is common to s_{14} and s_{54}.

Many simple techniques can be used for determining the elements of that sum $(s_{1i}+s_{4j}+s_{5k})$ whose implementation requires the minimum number of gates. In particular the procedure, which is usually used for the conventional covering problem, can be applied to the solution of this problem.

Let e_p be the binary variable which denotes the condition that p is implemented in the considered network; that is, $e_p = 1$ denotes that p is implemented in the minimal network and $e_p = 0$ denotes the contrary. The following condition must be verified:

$$\Big[\big(e_{\bar{b}}\cdot e_{\bar{c}+\bar{d}}+e_{\bar{a}+\bar{b}}\cdot e_{\bar{c}+\bar{d}}+e_{\bar{b}}\cdot e_{\bar{a}+\bar{c}+\bar{d}}+e_{\bar{a}+\bar{b}}\cdot e_{\bar{a}+\bar{c}+\bar{d}}\big)\cdot\big(e_{\bar{a}}\cdot e_{\bar{c}}+e_{\bar{a}+\bar{b}}\cdot e_{\bar{c}}+e_{\bar{a}}\cdot e_{\bar{b}+\bar{c}}+e_{\bar{a}+\bar{b}}\cdot e_{\bar{b}+\bar{c}}\big)\cdot\big(e_{\bar{b}}\cdot e_{\bar{a}+\bar{d}}+e_{\bar{b}+\bar{c}}\cdot e_{\bar{a}+\bar{d}}+e_{\bar{b}}\cdot e_{\bar{a}+\bar{c}+\bar{d}}+e_{\bar{b}+\bar{c}}\cdot e_{\bar{a}+\bar{c}+\bar{d}}\big)\Big]= 1$$

This expression can be easily expressed as a sum of products of variables, and every its prime implicant denotes an irredundant selection. The addendum which contains the minimum number of variables - in this case, $e_{\bar{a}+\bar{b}} \cdot e_{\bar{a}+\bar{c}+\bar{d}} \cdot e_{\bar{b}+\bar{c}}$ - corresponds to the minimal TANT expression. Figure 4 shows the minimal TANT network implementing the assigned function; it corresponds to expression (9).

ACKNOWLEDGEMENT.

The author wishes to express his gratitude to Professor Rinaldo Sartori and to Professor Giuseppe Biorci for their encouragement and guidance in the preparation of this paper.

VI. REFERENCES.

1. T. C. Bartee, "Design using computers", in A Survey of Switching Circuit Theory, E. J. McCluskey Jr. and T.C. Bartee, eds., McGraw-Hill Book Co., N.Y., 1962. Pp.89-102.
2. Gerald A. Maley and John Earle, The Logic Design of Transistor Digital Computers, W. L. Everitt, ed., Prentice - Hall, Englewood Cliffs, N.J.; 1963.
3. L. Dadda, "Le operazioni di Pierce e di Sheffer", Resoconti dell'Istituto Lombardo, 1959, Pp. 565.
4. E. J. McCluskey, "Logical Design Theory of NOR Gate Networks with No Complemented Inputs", Proc. of the Fourth Ann. Symp. on Switch. Circ. Th. and Log.Des., IEEE, Inc. 1963. Pp. 137-148.
5. W. V. Quine, "A way to simplify truth functions",Amer. Math. Monthly, vol. 62, no. 9, 1955. Pp. 627-631.

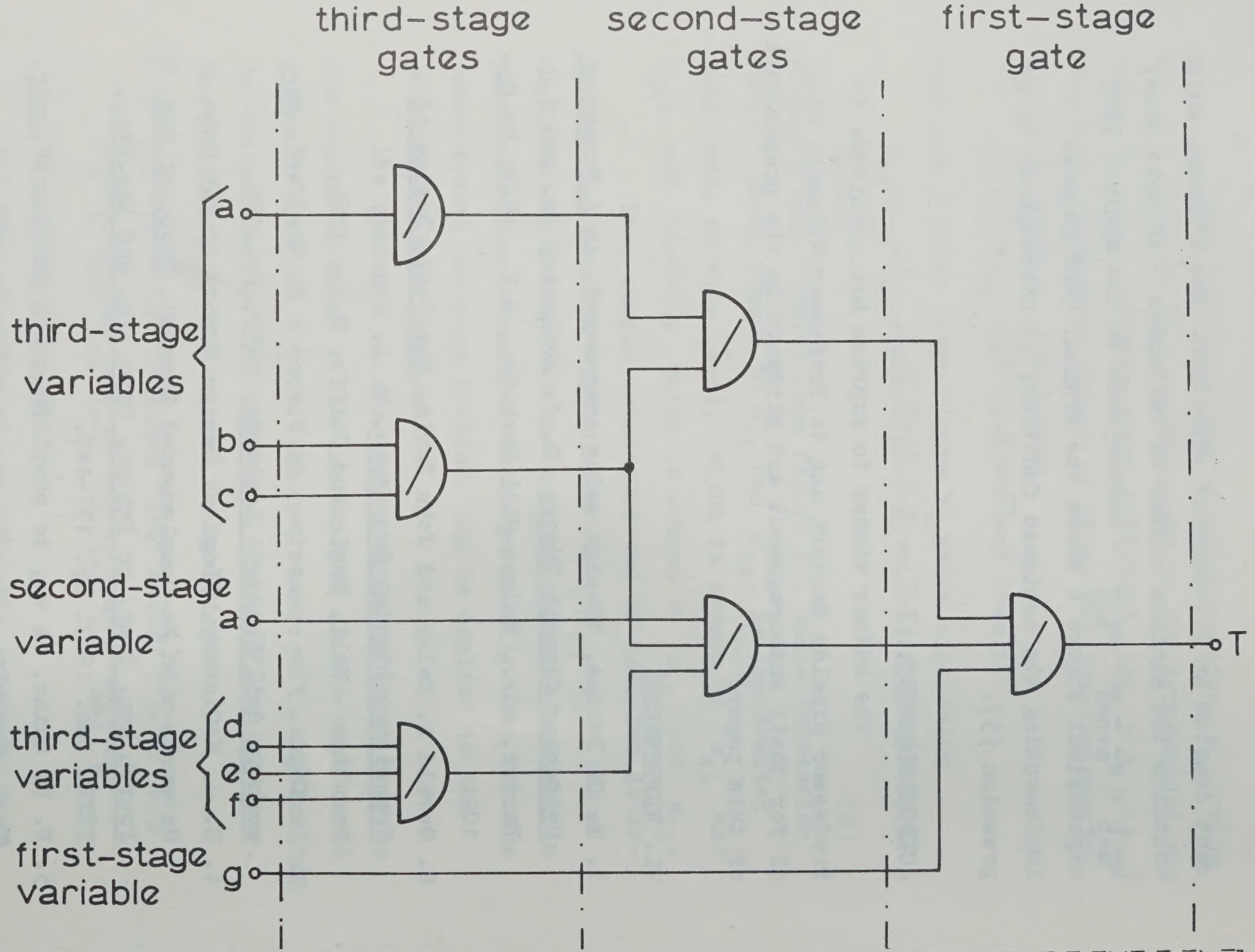

third-stage gates
second-stage gates
first-stage gate
a
third-stage variables
b
c
second-stage variable
a
third-stage variables
d
e
f
first-stage variable
g
T

Fig.1. General Structure of a TANT network. The function implemented is $T \sim [\overline{a} \cdot (\overline{b}+\overline{c})+\overline{a}(\overline{b}+\overline{c})(\overline{d}+\overline{e}+\overline{f})+\overline{g}]$.

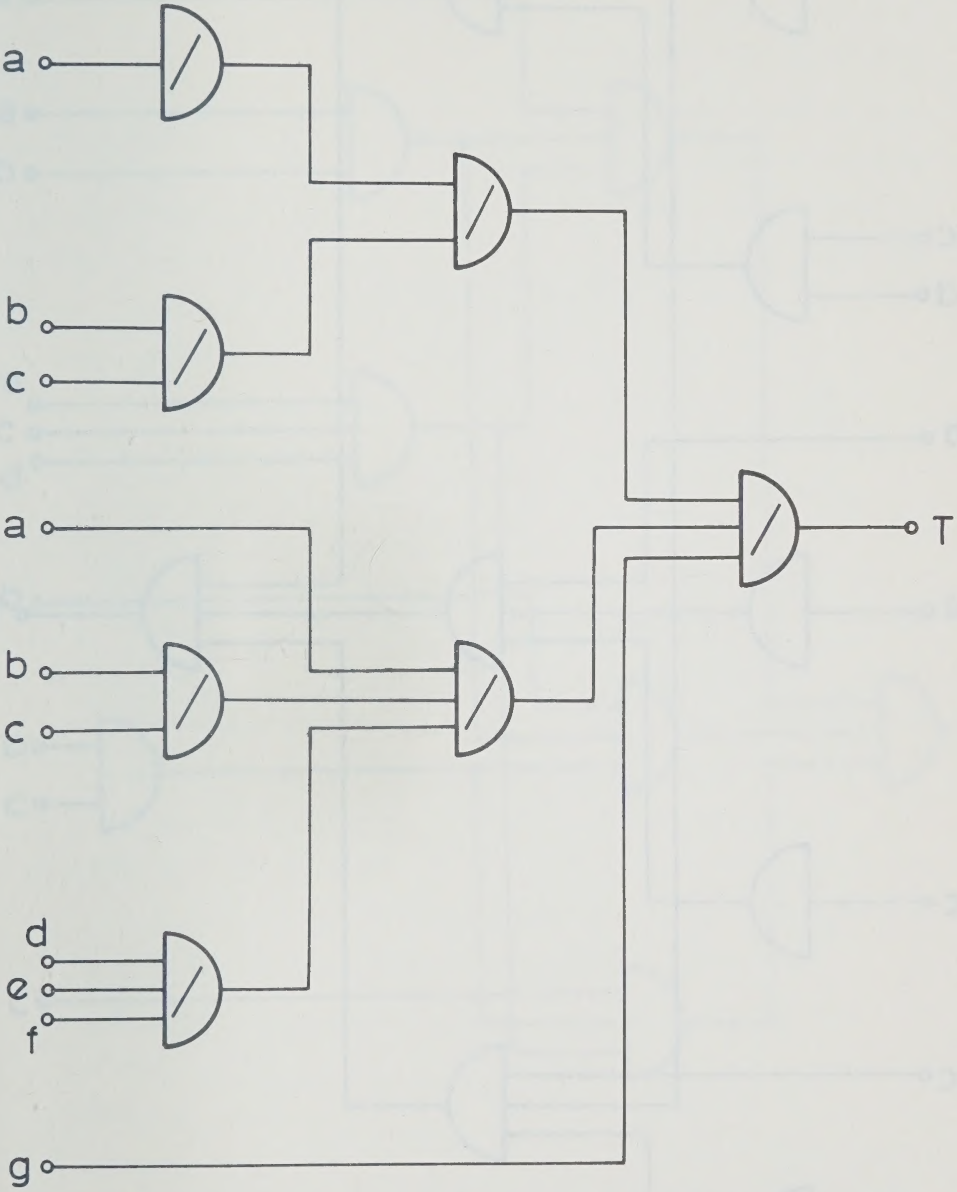


Fig. 2. Another example of TANT network implementing function $T \sim [\overline{a} \cdot (\overline{b}+\overline{c}) + a(\overline{b}+\overline{c})(\overline{d}+\overline{e}+\overline{f}) + \overline{g}]$ (see also figure 1).

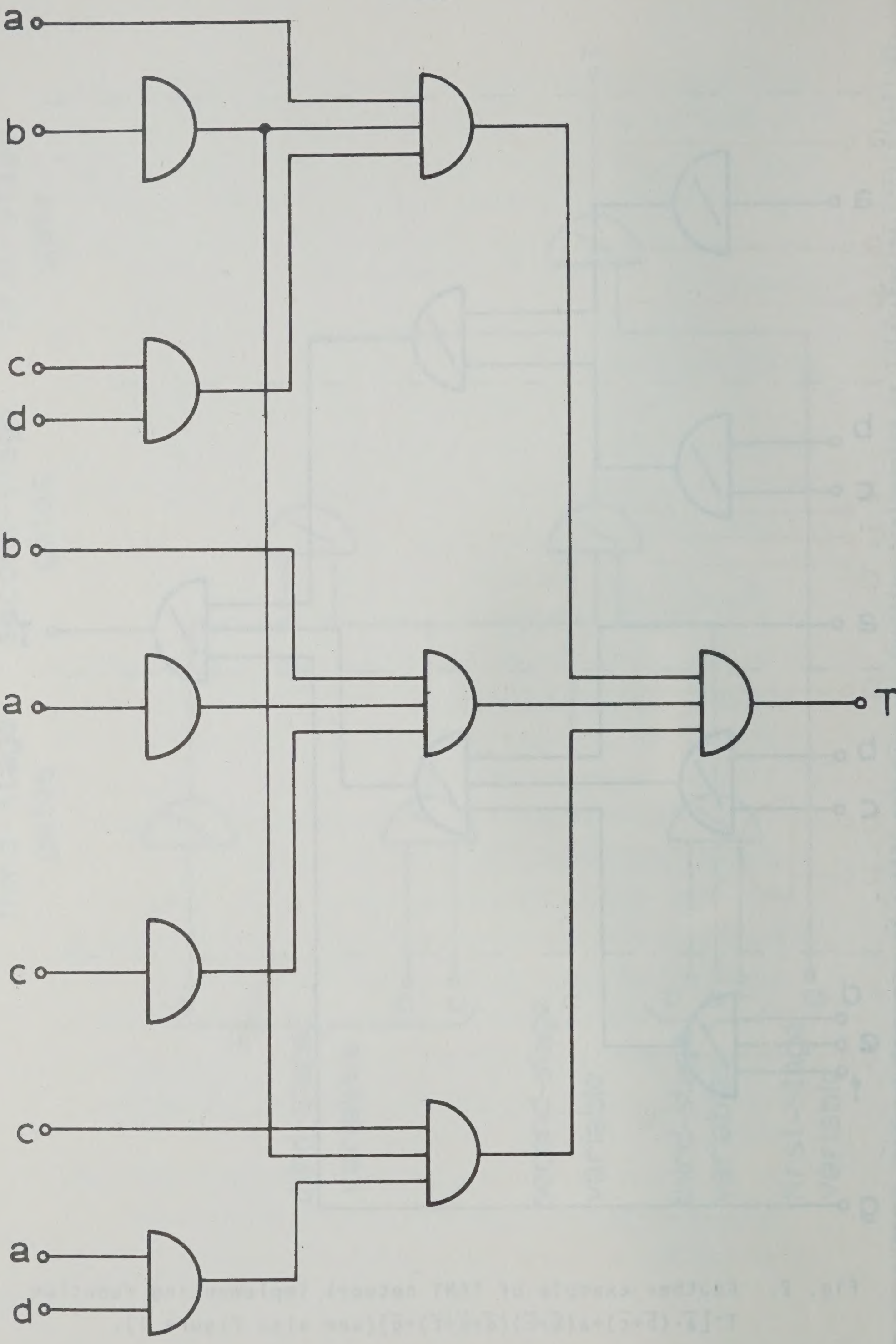


Fig. 3. Implementation of expression: $[a\overline{b}\cdot(\overline{c}+\overline{d})+b\overline{ac}+c\overline{b}\cdot(\overline{a}+\overline{d})]$

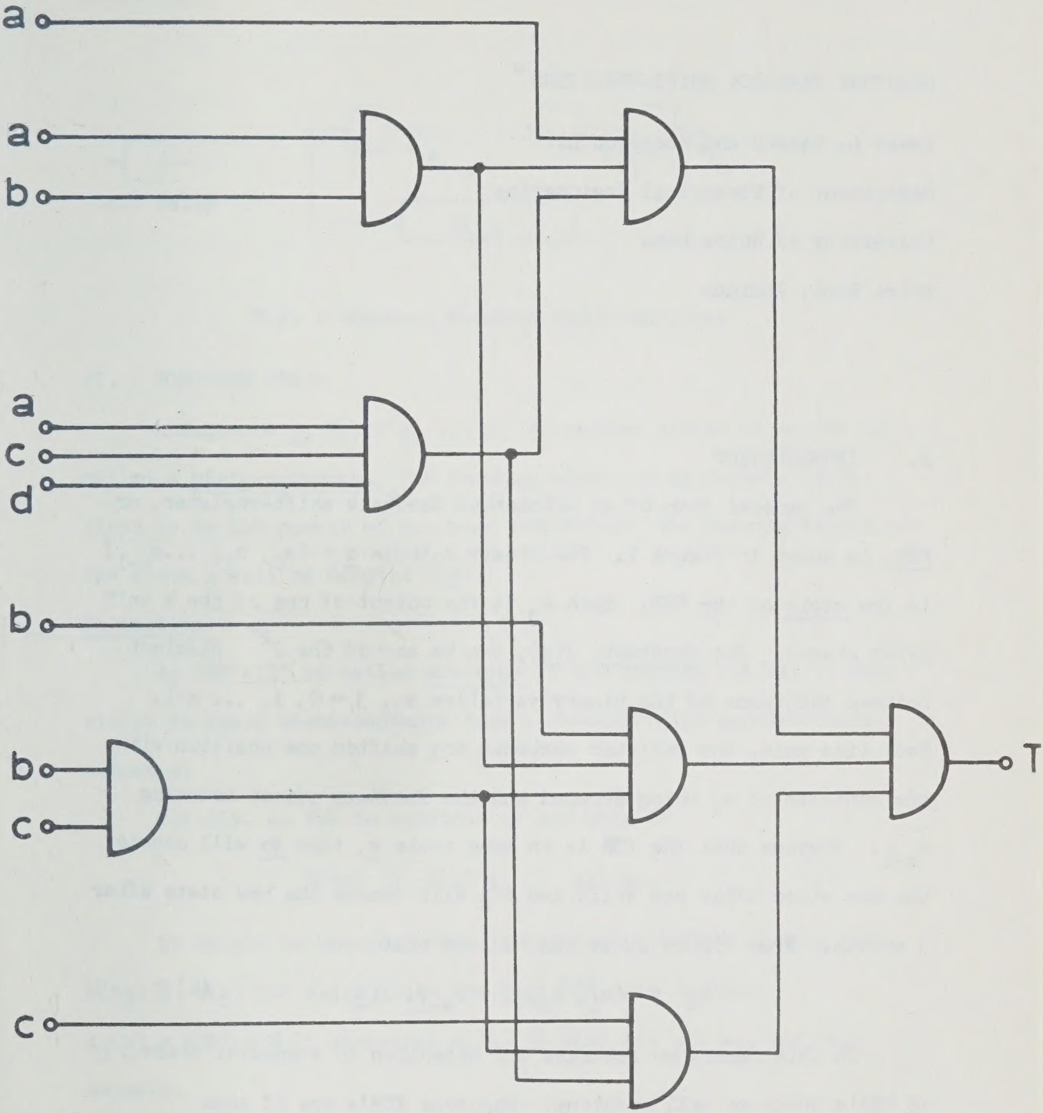


Fig. 4. Implementation of expression: $[a\cdot(\overline{a}+\overline{b})\cdot(\overline{a}+\overline{c}+\overline{d})+b(\overline{a}+\overline{b})\cdot(\overline{b}+\overline{c})+c\cdot(\overline{b}+\overline{c})\cdot(\overline{a}+\overline{c}+\overline{d})]$.

MONOTONE FEEDBACK SHIFT-REGISTERS*

James L. Massey and Ruey-wen Liu
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

I. INTRODUCTION

The general form of an autonomous feedback shift-register, or FSR, is shown in Figure 1. The binary m-tuple $\underline{s} = (s_0, s_1, \ldots s_{m-1})$ is the state of the FSR. Each s_j is the output of one of the m unit delay stages. The feedback, $f(\underline{s})$, can be any of the 2^{2^m} distinct boolean functions of the binary variables s_j, $j = 0, 1, \ldots m-1$. Each time unit, the register contents are shifted one position with the contents of s_0 being dropped and the feedback output becoming s_{m-1}. Suppose that the FSR is in some state $\underline{s}$, then $\underline{\theta s}$ will denote the new state after one shift and $\theta^j\underline{s}$ will denote the new state after j shifts. From Figure 1, it can be seen that

$$\theta\underline{s} = (s_1, s_2, \ldots s_{m-1}, f(\underline{s})). \qquad (1)$$

In this paper, we restrict our attention to a special class of FSR's which we call monotone. Monotone FSR's are of some

---

*This work was supported by the National Science Foundation under research grant GP-2547.

practical interest and will be shown to have interesting analytical properties.

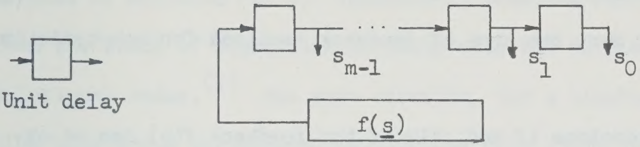


Fig. 1--General Feedback Shift-Register

II. MONOTONE FSR's

A sequence $\underline{s}$, $\theta\underline{s}$, $\theta^2\underline{s}$, ... of successive states of an FSR is called a state-sequence. The Hamming weight of an m-tuple is defined to be the number of non-zero components, the Hamming weight of the state $\underline{s}$ will be denoted $W(\underline{s})$.

Definition

An FSR will be called monotone if the Hamming weights of the states in every state-sequence form a monotonically non-increasing sequence.

Clearly, an FSR is monotone if and only if

$$W(\theta\underline{s}) \leqq W(\underline{s}) \ , \qquad \text{all } \underline{s}. \qquad (2)$$

It should be remarked that if the feedback f(s) is such that $W(\theta\underline{s}) \leqq W(\underline{s})$ for all $\underline{s}$, then the dual FSR with feedback $f_D(\underline{s}) = 1 \oplus f(\underline{s} \oplus \underline{1})$ satisfies $W(\theta\underline{s}) \geqq W(\underline{s})$ for all $\underline{s}$., and conversely.

(Note: $\oplus$ denotes addition modulo-two. $\underline{1}$ and $\underline{0}$ will be used to denote the m-tuples (1, 1, ... 1) and (0, 0, ... 0) respectively.) Moreover, any state-sequence of an FSR can be converted to a state-sequence in its dual by complementing all states. Thus the behavior of an FSR for which the Hamming weights of successive states form monotonically non-decreasing sequences can be determined from the behavior of its dual monotone FSR.

We consider next the type of feedback required for monotonicity.

Theorem 1

An FSR is monotone if and only if the feedback $f(\underline{s})$ can be expressed

$$f(s) = s_0\Phi(s_1, s_2, \dots s_{m-1}) \qquad (3)$$

where Φ is an arbitrary boolean function.

Proof

$f(s)$ can always be expanded on the single variable s_0 as

$$f(\underline{s}) = s_0\Phi(s_1, s_2, \dots s_{m-1}) \oplus (1 \oplus s_0)\beta(s_1, s_2 \dots s_{m-1}) \qquad (4)$$

where Φ and β are boolean functions. It remains to show that $\beta \equiv 0$ for monotonicity. From (1), we note that

$$W(\underline{s}) - W(\theta\underline{s}) = s_0 - f(s) \qquad (5)$$

and this quantity according to (2) must be nonnegative for a monotone FSR.

Suppose first that $\beta \equiv 0$. From (4) and (5), we have

$$W(s) - W(\theta s) = s_0\left[1 - \Phi(s_1, s_2, \dots s_{m-1})\right] \geqq 0$$

since Φ must be either 0 or 1. Hence the FSR is monotone.

Conversely, suppose $\beta(s_1, s_2, \dots s_{m-1}) = 1$ for $s_j = a_j$, $j = 1, 2 \dots m-1$. Consider the state $\underline{s}' = (0, a_1 \dots a_{m-1})$. From (4) and (5), we have $W(\underline{s}') - W(\theta\underline{s}') = -1$ so the FSR is not monotone.

Since Φ is a boolean function of m-1 variables, it follows immediately from theorem 1 that:

Corollary

There are exactly $2^{2^{m-1}}$ distinct m-stage monotone FSR's.

In the succeeding sections, we will study the general properties of monotone FSR's. Interest in monotone FSR's was generated from the study of sequential circuits embodied in decoders for convolutional codes.(1) Two such circuits, for a single-error-correcting code and a double-error-correcting code respectively, are shown in Figures 2(a) and 2(b).(2) The trivial circuit in Figure 2(a) is a monotone FSR. The circuit in Figure 2(b) is not an FSR, but an equivalent FSR can be found using a method developed by the authors.(3) The equivalent FSR is shown in Figure 3. Since $f(\underline{s}) = s_0(1 \oplus s_1 \oplus s_2)(s_3 \oplus s_4)$, it follows from theorem 1 that this is also a monotone FSR.

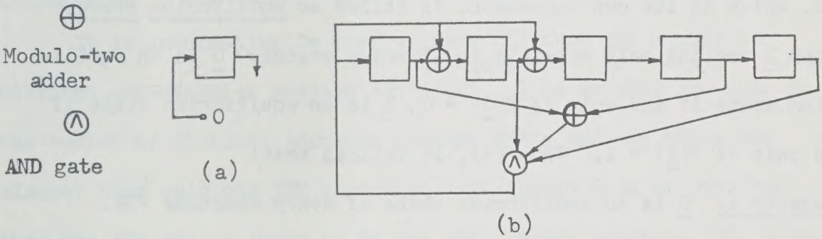
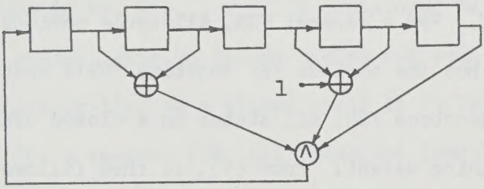


Fig. 2--Sequential Circuits in Convolutional Decoders



$$f(\underline{s}) = s_0(1 \oplus s_1 \oplus s_2)(s_3 \oplus s_4)$$

Fig. 3--Equivalent FSR for Circuit in Figure 2(b)

III. STATE-DIAGRAM STRUCTURE

The state-diagram of an FSR consists of a set of 2^m nodes, each labelled as a distinct state, interconnected by directed lines showing the transitions from states to their successors. As an example, the state-diagram of a 3-stage monotone FSR is shown in Figure 4.

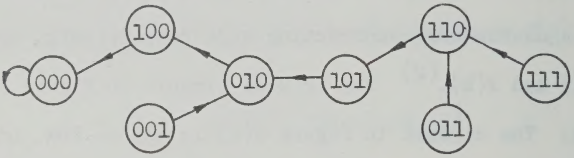


Fig. 4--State-Diagram of 3-Stage FSR, $f(\underline{s}) = s_0 s_1 (1 \oplus s_2)$

Each state in a state-diagram has a unique successor, but may have none, one, or two predecessors. A state $\underline{s}$ satisfying $\theta\underline{s} = \underline{s}$, i.e. which is its own successor, is called an equilibrium state. $\underline{0}$ and $\underline{1}$ are the only possible equilibrium states. $\underline{0}$ is an equilibrium state if and only if $f(\underline{0}) = 0$, $\underline{1}$ is an equilibrium state if and only if $f(\underline{1}) = 1$. From (3), it follows that:

Property 1: $\underline{0}$ is an equilibrium state of every monotone FSR.

A sequence of T distinct states $(T > 1)$ $\underline{s}_1, \underline{s}_2, \dots \underline{s}_T$ such that $\theta\underline{s}_T = \underline{s}_1$ and $\theta\underline{s}_j = \underline{s}_{j+1}$ $j = 1, 2, \dots T-1$ is called a closed cycle of period T. For a general FSR, all cycle periods up to 2^m are possible,(4) but the periods for monotone FSR's must be much smaller. For a monotone FSR, all states on a closed cycle must have the same Hamming weight. From (5), it then follows that $f(\underline{s}) = s_0$ for all states on the closed cycle. Thus the states on a closed cycle of a monotone FSR must be cyclic shifts of one another, and hence satisfy $\theta^m\underline{s} = \underline{s}$. Thus:

Property 2: The period T of any closed cycle of a monotone FSR is a divisor of the register-length m.

The pure-cycling register(4) or PCR is the FSR such that $f(\underline{s}) = s_0$ for all $\underline{s}$. From the preceeding remarks, it follows that any cycle in a monotone FSR is also a closed cycle in the m-stage PCR.

An FSR is said to be non-singular if every state is either an equilibrium state or lies on a closed cycle. Golomb and Welch (5) have shown that an FSR is non-singular if and only if $f(\underline{s})$ can be expressed as

$$f(\underline{s}) = s_0 \oplus \beta(s_1, s_2, \ldots s_{m-1}). \qquad (6)$$

Since $f(\underline{s}) = s_0$ for all states $\underline{s}$ of a monotone FSR which lie on closed cycles or are equilibrium states, it follows that:

Property 3: The only non-singular m-stage monotone FSR is the PCR.

It is interesting to observe from (6) that the number of distinct non-singular m-stage FSR's ($2^{2^{m-1}}$) is exactly the same as the number of distinct monotone m-stage FSR's and yet these two classes have only one FSR in common. It should also be remarked that the PCR can be shown to be the only linear monotone FSR except for the trivial register with $f(\underline{s}) \equiv 0$.

A sequence $\underline{s}, \theta\underline{s}, \theta^2\underline{s}, \ldots \theta^{L-1}\underline{s}$ containing no equilibrium states and no states lying on closed cycles and such that $\theta^L\underline{s}$ is an equilibrium state or lies on a closed cycle is called a transient of length L. For a general FSR, all transient lengths up to $2^m - 1$ are possible. However, the transients in a monotone FSR must be shorter.

Property 4: No transient in a monotone FSR can have length greater than $(m - 1)^2 + 1$.

To prove this property, consider the following transient formed by starting with $\underline{s} = \underline{1}$ and setting $f(s) = s_0$ for each state, except when a closed cycle would occur in a monotone FSR in which case setting $f(s) = 0$.

$$\begin{array}{ll}
(1, 1, 1, \ldots 1, 1) & \text{Weight } m \\
\left.\begin{array}{l}(1, 1, 1, \ldots 1, 0)\\ \vdots \\ (1, 0, 1, \ldots 1, 1)\end{array}\right\} & \text{Weight } m\text{-}1 \\
(0, 1, 1, \ldots 1, 0) & \\
\vdots & \\
(1, 0, 0, \ldots 0, 1) & \\
\left.\begin{array}{l}(0, 0, 0, \ldots 1, 0)\\ \vdots \\ (1, 0, 0, \ldots 0, 0)\end{array}\right\} & \text{Weight } 1
\end{array}$$

Clearly, this transient can occur in a monotone FSR. For each of the Hamming weights 1, 2, ... m-1, there are m - 1 states on the transient. Counting the state $\underline{1}$, we find a total of $(m-1)^2 + 1$ states on the transient. That no transient can be longer follows from the fact that if m consecutive states of any monotone FSR have the same Hamming weight, they must form a closed cycle.

The state-diagram in Figure 4 contains a maximal-length transient since $(m-1)^2 + 1 = 5$ for $m = 3$.

A branch-state is a state which has two distinct predecessors (one of which may be itself). For example, 110 and 000 are branch-states in the state-diagram of Figure 4.

Property 5: The number of branch-states of a monotone FSR is the number of distinct solutions of $\Phi(s_1, s_2, \ldots s_{m-1})=0$.

To prove this property, we note that if $\Phi(a_1, a_2, \ldots a_{m-1}) = 0$, then the state $(a_1, a_2, \ldots a_{m-1}, 0)$ has both $(0, a_1, \ldots a_{m-1})$ and $(1, a_1, \ldots a_{m-1})$ as predecessors. Conversely, if $(b_1, b_2, \ldots b_m)$ is a branch-state, then its predecessors must be $(1, b_1, \ldots b_{m-1})$ and $(0, b_1, \ldots b_{m-1})$. But the latter predecessor will have Hamming weight less than its successor unless $b_m = 0$. Hence, if the FSR is monotone, we must have $f(1, b_1, \ldots b_{m-1}) = \Phi(b_1, b_2, \ldots b_{m-1}) = 0$.

IV. STABILITY

One of the most interesting unexplored questionsin connection with FSR's is that of stability.$^{(1)}$ An FSR is said to be stable if $\theta\, \underline{0} = \underline{0}$ and for every state $\underline{s}$, there is some N such that $\theta^N \underline{s} = \underline{0}$. Thus $\underline{0}$ is an equilibrium state of a stable FSR and all other states lie on transients. For example, the FSR diagrammed in Figure 4 is stable.

It seems natural to begin the study of stability of FSR's with the class of monotone FSR's since these registers have the property that (considering Hamming distance as the metric) successive states can move toward the origin $\underline{0}$ but can never move farther away. In spite of such an obvious simplification of the analysis, we have been unable to derive a simple stability test for an arbitrary monotone FSR. The strongest result we have thus far obtained is:

Theorem 2

A monotone FSR for which the feedback has the form

$$f(\underline{s}) = s_0 s_k \beta(s_1, \ldots s_{k-1}, s_{k+1}, \ldots s_{m-1}) \quad (7)$$

where k and m are relatively prime, is stable if $\beta(1, 1, \ldots 1) = 0$ and is unstable if $\beta(1, 1, \ldots 1) = 1$.

Proof

If $\beta(1, 1, \ldots 1)$, then from (7) $f(\underline{1}) = 1$ so that $\underline{1}$ is an equilibrium state of the FSR and the FSR is unstable.

Suppose then that $\beta(1, 1, \ldots 1) = 0$ so that $\underline{1}$ is not an equilibrium state. It remains to show that the FSR then has no closed cycles. Assume on the contrary that $(a_0, a_1, \ldots a_{m-1})$ is a state of the FSR lying on a closed cycle. Consider then the doubly-infinite sequence generated from this state

$$\ldots a_{-3}\, a_{-2}\, a_{-1}\, a_0 \ldots a_{m-1}\, a_m \ldots \qquad (8)$$

by the rule $a_{j + nm} = a_j$ for all n and for all j such that $0 \leqq j \leqq m-1$. Since a closed cycle of a monotone FSR is also a cycle in the PCR, every m consecutive digits in (8) is a state on the same closed cycle.

If $s_0 = 1$ and $s_k = 0$, then from (7) $f(\underline{s}) = 0$ and according to (5) $\theta\underline{s}$ and $\underline{s}$ do not have the same Hamming weight so that $\underline{s}$ cannot lie on a closed cycle. Hence, if $a_j = 1$ in (8), then it must be that $a_{j + ik} = 1$ for all $i \geqq 0$.

At least one of the digits in (8) must be a "one". We now show that all the digits in (8) must then be "ones" which will contradict the original assumption since the state (1, 1, ... 1) cannot lie on a closed cycle of a monotone FSR. Suppose $a_j = 1$. From the previous remarks, it follows that $a_{j+ik+nm} = 1$ for all n and all $i \geqq 0$. By the hypothesis, the greatest common divisor of k and m is unity. By the well-known property of the greatest common divisor, there exist integers I and J with $I \geqq 0$ such that

$$1 = Ik + Jm.$$

Let a_{j+h} be an arbitrary digit in (8). It follows then that

$$a_{j+h} = a_{j+hIk+hJm} = a_j = 1$$

as was to be shown.

From theorem 2, it follows that if k = 1 and $\beta(1, 1, \ldots 1)=0$ then the monotone FSR with feedback as in (7) is stable. The value of β is arbitrary for its remaining $2^{m-2} - 1$ distinct arguments. Hence:

Corollary

There are at least $2^{2^{m-2}-1}$ distinct stable m-stage monotone FSR's.

This corollary indicates that stable monotone FSR's are quite numerous, numbering at least half the total of all monotone FSR's with one fewer stage.

The FSR in Figure 2(a) is trivially stable according to theorem 2 since $\beta \equiv 0$. Less trivially, the stability of the FSR diagrammed in Figure 4 is also predicted from theorem 2. However, theorem 2 is not strong enough to show the stability or instability of the FSR in figure 3. This FSR can be shown to be stable by working out its complete state-diagram.

It should be remarked that the stability of the FSR's embodied in convolutional decoders is a matter of practical import. If the FSR is not stable, then a decoding failure may lead to an indefinite number of further decoding failures.(1)

V. SUMMARY

Monotone FSR's have been introduced and an extensive characterization of their state-diagrams has been given. This class of

FSR's is proposed as a starting point in the stability analysis of FSR's. The beginning of such a theory was presented, but the formulation of a simple and general test for stability of monotone FSR's remains an unsolved problem. It is hoped that this paper will stimulate other workers to join the attack on this problem.

VI. References

1. J. L. Massey and R. Liu, "Application of Lyapunov's Direct Method to the Error-Propagation Effect in Convolutional Codes" IEEE Trans. Inf. Th., IT-10, July, 1964. Pp.248-250

2. J. L. Massey, Threshold Decoding, M.I.T. Press, Cambridge, Mass., 1963. Pp.42-55

3. J. L. Massey and R. Liu, "Equivalence of Nonlinear-Feedback Shift-Registers," to appear in IEEE Trans. Inf. Th.

4. S. W. Golomb, L. R. Welch, and R. M. Goldstein, "Cycles from Nonlinear Shift Registers," JPL Prog. Rpt. No. 20-389, Calif. Inst. of Tech., Aug. 31, 1959.

5. S. W. Golomb and L. R. Welch, "Non-linear Shift Register Sequences," JPL Memo. 20-149, Calif. Inst. of Tech., Oct. 25, 1957.

GENERALIZED PROCEDURES FOR THE DETERMINATION OF GENERATING FUNCTIONS OF DIRECTED GRAPHS OF DISCRETE SYSTEMS

C. V. RAMAMOORTHY
HONEYWELL INC
NEEDHAM, MASS.

1. DISCRETE SYSTEM REPRESENTATION BY DIRECTED GRAPHS

A directed graph is a set of nodes connected by directed branches and such that a complex number called the branch transmission is attached to each branch. Thus a graph of n nodes can be represented by an n x n branch transmission matrix T, which exhibits the relationships between the nodes. This matrix is unique to the graph upto a permutation of entries due to the renumbering of the nodes. In this paper, we call t_{ij} as the branch transmission from node i to node j and restrict our discussion to linear graphs which are governed by two rules. viz.,

a. Addition rule, by which parallel identically directed branch transmission between any two nodes can be replaced by their sum in the same direction and

b. Multiplication rule, by which branch transmissions in series between two nodes can be represented by a transmission which is equal to the product of the original branch transmissions (1,2,3,).

It can be shown(4) that this formulation allows a unified representative for path enumerations, transfer functions and discrete probabilistic (Markov) systems.

Given an abstract graph, whether representing sequential processes, signal flows or connectivities, there arises a need to know the total transmission from one set of nodes to another set as a function of number of branches traversed. Or, from the path enumeration viewpoint, it is required to know the number of different ways of reaching from one set of nodes to another set as a function of number of branches traversed in series. Such a transmission (e.g., path enumeration) function is called the generating function between the given set of nodes (terminal nodes) and is of considerable interest in the study of communication and transport networks, information retrieval and biophysics. (2,3,4) The generating function is a storehouse of information about the given graph and one can get from it many interesting properties of the graph by certain operations.$^{(4)}$ (Figure 5)

2. DERIVATION OF THE GENERATING FUNCTION

Consider an abstract graph of n nodes whose branch transmission matrix is given by $\underline{T}$. Let $\underline{T}^n = \underbrace{\underline{T}.\underline{T}.\underline{T}.\dots.\underline{T}}_{n\text{ terms}}$ with $\underline{T}^o = \underline{I}$, the identity matrix. The transmission from an arbitrary node i to another node k traversing only j branches in series is given by $g_j^{ik} = (\underline{e}^i) \left[\underline{T}^j\right]_{ik} (\underline{e}_k)$ where $(\underline{e}^i)$ is a row vector of n elements such that its p-th element is

$$
\begin{aligned}
(\underline{e}^i)_p &= 1 \quad \text{if} \quad i = p \\
&= 0 \quad \text{otherwise}
\end{aligned}
$$

and $(\underline{e}_k)$ is a column vector of n elements such that its p-th element is

$$
\begin{aligned}
(\underline{e}_k)_p &= 1 \quad \text{if} \quad k = p \\
&= 0 \quad \text{otherwise}
\end{aligned}
$$

The generating function from node i to node k is defined by $G_{ik}(z) = \sum_{j=0}^{\infty} g_j^{ik} \; z^j$ in which z is a complex valued indepenent variable such that $|z| < \frac{1}{M}$ where M is the largest row sum of $\underline{T}$. $G_{ik}(z)$ is well defined and is equal to

$$G_{ik}(z) = \underline{e}^i \; \left[\underline{I}\text{-}\underline{T}z\right]^{-1} \; \underline{e}_k = \frac{A_{ki}}{\underline{A}} \qquad (2.1)$$

where $\underline{A} = \left[\underline{I}\text{-}Tz\right]$ and A_{ki} is the co-factor of ki-th element of $\underline{A}$ and $|\underline{A}|$, its determinant. It can also be interpreted as a solution of a set of simultaneous equations. Thus, if $X_i \, . \, \underline{e}_i$ is a column vector whose i-th component is X_i and all others zero, and if $\left[\underline{B}\right] \underline{x} = \left[X_i\right] \underline{e}_i$ is a set of linear equations where $\underline{B} = \left[\underline{I}\text{-}\underline{T}^T{}_z\right]$ then $G_{ik}(z) = \dfrac{x_k}{X_i}$. (2.2)

Thus, the n components of $\underline{x}$ can be interpreted as node signals of an appropriately modified signal flow graph$^{(1,4)}$, and $G_{ik}(z)$ as the signal appearing at node k per unit external signal X_i injected at node i.

Applying Cramer's rule in Equation 2,1 we get $G_{ik}(z) = \frac{x_k}{X_i} =$

$$\frac{\begin{vmatrix} (1-t_{11}z) & -t_{21}z & -t_{k-1,1}z & 0 & -t_{n1}z \\ -t_{12}z & (1-t_{22}z) & -t_{k-1,2}z & 0 & -t_{n2}z \\ \hdashline -t_{1,i}z & -t_{2,i}z & -t_{k-1,i}z & 1 & -t_{ni}z \\ -t_{1n}z & -t_{2n}z & -t_{k-1,n}z & 0 & (1-t_{nn}z) \end{vmatrix}}{\begin{vmatrix} (1-t_{11}z) & -t_{21}z & -t_{k-1,1}z & -t_{k1}z & -t_{n1}z \\ -t_{12}z) & (1-t_{22}z) & -t_{k-1,2}z & -t_{k2}z & -t_{n2}z \\ \hdashline -t_{1i}z & -t_{2i}z & -t_{k-1,i}z & -t_{ki}z & -t_{ni}z \\ -t_{1n}z & -t_{2n}z & -t_{k-1,n}z & -t_{kn}z & (1-t_{nn}z) \end{vmatrix}} \quad (2.3)$$

3. TOPOLOGICAL METHOD OF DERIVING THE GENERATING FUNCTION

We shall next show that $G_{ik}(z)$ can be obtained by considering only the loops in the given graph and its equivalent numerator graph. For this, let us define a loop as a sequence of distinct nodes 1, 2, .. j, 1 such that the product $t_{12} \cdot t_{21} \cdots t_{j1}$ is not zero.

A self-loop is a loop with exactly one node. A non-touching (non-intersecting) loop set is a set of loops not having any node or branch in common.

4. DETERMINATION OF THE DENOMINATOR DETERMINANT

The determinant of the deonominator in (2.3) is known as the characteristic function of the graph. It is given by

$$|D\ (z)| = 1 - \sum P_{a_1} + \sum P_{a_2} - \sum P_{a_3} \cdot \cdot$$

where P_{a_m} is the product of transmittances of the a-th set of m non-touching loops in the modified graph; the summation is taken over all possible non-touching loops when taken m at a time. The proof is obtained directly by the application of Mason's loop rule for the denominator of a signal flow graph, when the latter is appropriately modified.(4)

5. DETERMINATION OF THE NUMERATOR DETERMINANT

Every determinant can be considered as the characteristic function of some graph. Thus given an arbitrary n-th order determinant, one can find an n-node graph whose characteristic function is the given determinant. We shall call this the equivalent graph of the given determinant.

Consider the numerator determinant in (2.3). This determinant can be mapped to an equivalent graph whose transmission matrix T' is given by

$$
\begin{aligned}
t'_{pq}z &= t_{pq}z && \text{if } p \neq k \\
&= 0 && \text{if } p = k \text{ and } q \neq i,k \\
&= -1 && \text{if } p = k \text{ and } q = i \\
&= 1 && \text{if } p = k \text{ and } q = k
\end{aligned}
$$

where t'_{pq} and t_{pq} are corresponding elements of the numerator and denominator graph transmission matrices. Thus, the numerator determinant is the characteristic function of the graph whose transmission matrix is $\underline{T}'$. Thus we have the following result: the generating function $G\ \ (z)$ of a graph whose transmission matrix is $\underline{T}$ is the ratio of the characteristic functions of the graph whose transmission matrices $\underline{T}'$ and $\underline{T}$ respectively defined by the above relation. This result is equivalent to Mason's gain rule(1) when applied on a modified graph. The usefulness of this expression lies in the uniform treatment of both the numerator and denominator expressions, requiring no consideration of paths and loop-and-path intersections.

Thus, for example, the path enumerating generating function from an arbitrary node i to another node k is given by the transmission function corresponding to those nodes when the given graph is treated as a signal flow graph after each branch transmission of the original graph is multiplied by z which is a seperator variable. The generating function of a directed graph from a set of starting nodes S to a set of end nodes E is given by $G_{S,E}(z) = \sum_m \sum_n G_{m \in S,\ n \in E}(z)$, since this is a linear process where the principle of superposition applies.

6. CONNECTIVITY MATRIX

In many instances as in the characterization of the necessary and sufficient conditions of generating functions(4), the actual value of branch transmission t_{ij} is not important but merely whether it is non-zero. Thus, $\underline{C}$ is defined as the connectivity matrix of a graph whose branch transmission matrix is $\underline{T}$

$$
\begin{aligned}
\text{if } c_{ij} &= 1 \quad \text{if } |t_{ij}| > 0 \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

We shall briefly consider the problems of deriving the generating functions of large graphs. We will be concerned with the reduction of complexity of the graph rather than the required precision in numerical calculation. Hence we shall work with the connectivity matrix rather than the transmission matrix.

It had been shown earlier that one can derive the generating function using either the topological approach or the algebraic approach requiring the inversion of a matrix (Section 2). The decision which method is best will depend on a) the size of the graph in terms of the number of nodes and b) the topological complexity as indicated by the number of loops in it. We shall classify all graphs from a computational view point as a) small graphs b) large topologically well-ordered graphs and c) large random graphs. In a well-ordered graph, it is easy to find loops, without regard to the number of nodes. All other graphs are considered random.

7. SMALL GRAPHS OR LARGE WELL-ORDERED GRAPHS

When the loops follow a pre-determined pattern with the addition of each node or when they are evident by inspection, the topological method is better than matrix method. *

8. LARGE RANDOM GRAPHS

If no topological ordering is evident in the graph, we have to decide the best method or combination of methods that should be used for this problem. We propose the following procedure for this.

STEP 1: Using the connectivity matrix, remove all the redundant nodes (i.e., those whose deletion does not change the generating function) by applying the irredundant node algorithm(4).

STEP 2: Apply on the reduced graph, the algorithm that partitions the nodes into strongly connected subgraphs(4).

STEP 3: We next estimate the number of distinct loops possible in each strongly connected subgraph using the algorithm in the next section. We do this for all strongly connected subgraphs and estimate the total number of loops possible in the graph.

If the number of loops are excessive, iterative node elimination method (Section 10) is recommended.

---

*The loops can be found from connectivity matrix by algorithms in Reference 4.

9. TEST TO DETERMINE THE COMPLEXITY OF THE GRAPH

For the purpose of determination of the generating function, the complexity of a random graph can be measured by the number of loops in it, after the removal of all of its redundant nodes. We take each strongly connected subgraph and estimate the number of loops in it as follows.

First we look at the connectivity matrix of the subgraph and determine the number of self-loops which is the trace of the matrix. Next, we note if n_i is the number of nodes in the subgraph, its closeness to the complete graph (maximum number of loops) can be measured by summing its branching ratios (number of branches emanating from each node) and comparing it to n_i^2.

A complete graph of n-nodes has a connectivity matrix of all ones. The number of distinct directed loops in it is $\sum_{i}^{n} \frac{n!}{(n-i)!\,i}$

If R_i is the sum of the branch ratios of the i-th non-intersecting strongly connected subgraph, then one can use the ratio

$$\left[\frac{R_i}{n_i^2}\right]^{\rho} \times \left[\sum_{k=1}^{n_i} \frac{n_i!}{(n_i-k)!\,k}\right] \quad \text{as}$$

an estimate of the number of loops in it where $\rho = \frac{n_i^2}{R_i}$. $\rho = 2$ appears to be a good choice in many cases.

Another estimate is the cyclomatic number of Bergé viz., $(R_i - n_i + 1)$, which gives the number of distinct loops possible in a non-directed graph$^{(4)}$. Thus, for the whole graph the estimate for the total number of loops will be $\sum_i (R_i - n_i + 1)$ where the summation is over all the strongly connected maximal subgraphs. Based on the personal experience of the writer, the topological method becomes unwieldy when the ratio of loops to nodes exceeds 2.

10. A NODE REDUCTION PROCEDURE FOR LARGE RANDOM GRAPHS

Motivation: We assume in this section that the test for complexity has been performed on a given graph and the loops in it are found to be too many for application of the topological method.

We first derive the connectivity matrix from the given transmission matrix. Our motive will be to eliminate all non-terminal nodes one at a time. We shall choose the node to be removed by looking at the appropriate connectivity matrix at each stage of reduction and evaluating the amount of computation required to remove it. Generally, we shall choose those nodes whose cost of removal is minimum at that stage of computation.

11. CONVENTIONAL METHOD OF NODE REMOVAL

Since our proposed method is refinement of conventional method of node removal(1,4) we shall briefly describe the latter.

Consider the correspondence between the graph and a set of simultaneous equations. Let r be a non-terminal node (Fig. 1), and its removal is achieved as follows:

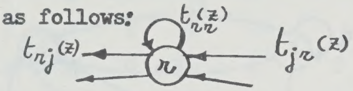


FIG. 1 - A NODE AND ITS BRANCHES

The equation corresponding to this node is

$$x_r = \sum_j t_{jr} x_j = \frac{\sum_{j \neq r} t_{jr} x_j}{1 - t_{rr}}$$

Consider now another node k, then,

$$x_k = \sum_j t_{jk} x_j = \sum_{j \neq r} \left[t_{jk} x_j + \frac{t_{rk} t_{jr} x_j}{1 - t_{rr}} \right]$$

One can thus remove node r by changing (and some times adding) some branch transmissions between other nodes.

12. STEPS IN THE PROPOSED NODE REDUCTION PROCEDURE

1. Reduce the given graph by removing all the redundant nodes.
2. Derive from the connectivity matrix of the reduced graph, and the node weights as follows: The node weight of node i is the product of the sums of one's in the i-th row including any "one" on the main diagonal and the sums of one's in the i-th column excluding any "one" on the main diagonal. Attach node weight to each non-terminal node of the reduced graph.
3. Pick the non-terminal nodes with the lowest node weight. Find their cost of elimination. If the node weight is w_i, then its removal requires a total of w_i multiplications and divisions and less than w_i additions. If there are two nodes with same minimum node weight, we choose the one that adds

least number of new branches in the new graph. We shall give a procedure to find this (Sec. 13).

4. We eliminate the node thus selected by appropriately modifying the graph and changing some branch transmissions (Sec. 13).
5. We go back to Step 2 and iterate the procedure until only the terminal nodes remain. Their branch transmissions are the components of their generating function.

Example -

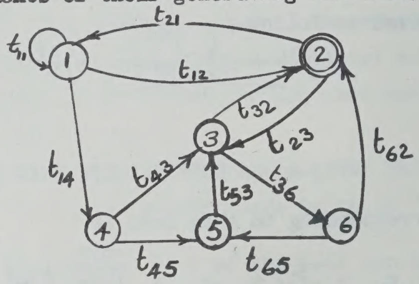


FIG. 2 - A GIVEN GRAPH

Starting node = 1
End node = 2
Connectivity matrix

	1	2	3	4	5	6	Node Wt.
1	1	1	0	1	0	0	
2	1	0	1	0	0	0	
3	1	1	0	0	0	1	9
4	0	0	1	0	1	0	2
5	0	0	1	0	0	0	2
6	0	1	0	0	1	0	2

Non-terminal nodes with least node weights are 4, 5, and 6. Consider node 4. Number of changes in branch transmission due to its removal is given by the intersection of its row "one" with its column "ones", in its connectivity matrix (Sec. 13), viz, t_{13} and t_{15}, both of which are new branches. Consider node 5. Its removal changes t_{43} but adds t_{63}. Similarly for node 6, we note t_{32} is changed and t_{35} is added.

Let us choose node 5 (or 6) since its removal requires least number of branch additions and changes.

The new branch transmission $t_{63}^{(1)}$ that is to be added will be $= t_{63}^{(1)} = t_{65}.t_{53}$. The branch transmission from node 4 to 3 will be $t_{43} + t_{45}.t_{53} = t_{43}^{(1)}$.

Thus, the new connectivity matrix will be:

$$\begin{array}{cccccccc} & 1 & 2 & 3 & 4 & 6 & \text{Node Wt.} \\ 1 & 1 & 1 & 0 & 1 & 0 & \\ 2 & 1 & 0 & 1 & 0 & 0 & \\ 3 & 1 & 1 & 0 & 0 & 1 & 9 \\ 4 & 0 & 0 & 1 & 0 & 0 & 1 \\ 6 & 0 & 1 & 1 & 0 & 0 & 2 \end{array}$$

This corresponds to the graph is Fig. 3

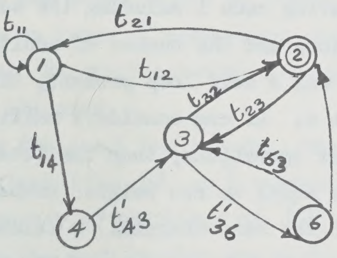


FIGURE 3

By iterating the procedure, we next select node 4, node 6 and node 3 consecutively until the graph is reduced to a two-node graph with node 1 as the input node and node 2 as the end node, with the generating function as

$$G_{12} = \frac{t_{12}^{(4)}}{1 - \left[t_{11}^{(4)} + t_{22}^{(4)} + t_{12}^{(4)}\ t_{21}^{(4)}\right] + t_{11}^{(4)}\ t_{22}^{(4)}}$$

where the superscripts on t's indicate they are the result of the 4th iteration. Note that all t's here are functions of z.

13. PROCEDURE TO SELECT NODES FOR ELIMINATION

Removing a non-terminal node i involves the following computations -

a. If node has a self loop t_{ii}, then we divide all incoming branch transmission by $1-t_{ii}$, i.e., $t_{ji}^{(s)} = \frac{t_{ji}}{1-t_{ii}}$ for all $j \neq i$ (cf. Section 11) and proceed to the next step.

b. For all nodes j and k not equal to i

$$t_{jk}^{(new)} = t_{ji}\, t_{ik} + t_{jk}^{(previous)}$$

This implies then i) the number of branch transmissions to be changed will be the product of the number of branches entering node i and the number of branches leaving it excluding the self-loop and ii) some new branches will be added corresponding to those $t_{jk}^{(previous)}$'s which were zero. If c is the number of branches entering and r is the number of branches leaving node i exluding the self loop, then the number of multiplications and the number of additions will both be equal to rc. If there was a self loop present, then the number of divisions will be equal to c. If one considers multiplication and division as equally complex operations, then the total number of complex operations will be equal to row weight (number of 1's in the 1-th row including that in the main diagonal) x column weight (number of 1's in the i-th column excluding any ones on the main diagonal) when taken from the connectivity matrix. This is precisely our node weight. Since we select the node with the least node weight we remove the one which is computationally least expensive. When we have to choose between two nodes which have the same minimum node weight, we will require that we select the one that adds the least number of new branches.

This process of selection is simplified if we note that the branch transmissions which will be changed are those that lie on the intersection of "one's" in the i-th column with the "one's" in the i-th row in the connectivity matrix. Thus, for example (cf. Fig. 4)

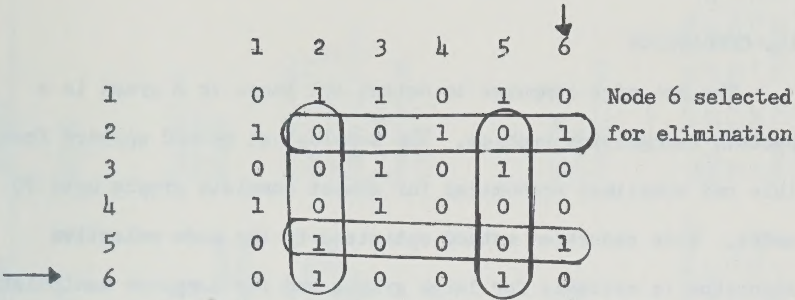


FIG. 4 - CHANGES IN BRANCH TRANSMISSION DUE TO ELIMINATION OF NODE 6.

if node 6 is selected for elimination, the branch transmission t_{22}, t_{25}, and t_{55} (corresponding to zero connectivities of Fig. 4) will be new branches and t_{52} will be changed. This explains our purpose of using the connectivity matrix.

14. NUMBER OF COMPUTATIONS

We shall use the total number of computations (multiplications and divisions) as an index of comparison against standard numerical matrix procedures. We shall consider a complete graph whose connectivity matrix is all ones.

If initially the graph had n nodes, the reduction of the graph to two nodes requires $1/3 \left[n^3 - n \right]$ computations. Incidentally, we note that total number of additions is $\left[\frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6} \right]$. Since a complete graph is the most complex graph one could find, the number of computations for an arbitrary graph will be less. We note that the Jordan-Gauss elimination procedure requires $\frac{n^3}{3}$ multiplications and $\frac{n^3}{3}$ additions, thus our method appears comparable to the best numerical procedures.

This method of node reduction based on minimizing the cost at each stage with a connectivity matrix is applicable to other areas like network analysis, and network flow problems, where node weights are computed differently.

15. CONCLUSION

The use of a computer to detect all loops in a graph is a pattern recognition problem. The topological method appears feasible and sometimes economical for almost complete graphs upto 20 nodes. Node reduction method optimized by the node selection algorithm is suitable for large graphs and for computer manipulation.

ACKNOWLEDGEMENT

The writer acknowledges with gratitude the help, advice and suggestions of Prof. D.W. Tufts of Harvard University and encouragement of Drs. R.B. Lawrance and B. Rabinovici of Honeywell organization.

REFERENCES

1. S. J. Mason, "Further Properties of Flowgraphs", Proc. IRE (July 1959).

2. S. Seshu and M. B. Reed, Linear Graphs and Electrical Networks, Addison-Wesley Publishing Co., Inc. 1961.

3. W. H. Kim and R. T. Chien, Topological Analysis and Synthesis of Communication Networks, Columbia University Press. 1962.

4. C. V. Ramamoorthy, Ph.D Thesis, Harvard University, May 1964.

CHART OF EQUIVALENT CHARACTERISTICS(4)

DIRECTED GRAPHS	SIMULTANEOUS LINEAR EQUATIONS	PATH ENUMERATION	CODE GENERATION	SAMPLED-DATA **SYSTEMS**	SHIFT REGISTER SEQUENCES	DISCRETE MARKOV PROCESS
NODE	VARIABLE	NODE (PT OF INTERSECTION OF PATHS)	SYMBOL GENERATOR	POINT OF SIGNAL SUMMATION	ADDER	STATE
BRANCH	RELATION BETWEEN VARIABLES	UNITS OF PATH		CONNECTION BETWEEN COMPONENTS	PHYSICAL MEANS OF TRANSFERRING INFORMATION	TRANSITION BETWEEN STATES
BRANCH TRANSMISSION FROM i TO j	COEFFICIENT OF VARIABLE	NUMBER OF DIFFERENT WAYS OF REACHING FROM i TO j	NUMBER OF SYMBOLS GENERATED	TRANSFER FUNCTION OF COMPONENT	INDEX OF DELAYS AND PARALLELISM OF INPUTS	CONDITIONAL PROBABILITY OF SYSTEM OF BEING IN STATE j AFTER STARTING ON STATE i
z	A SEPERATOR VARIABLE (PART OF COEFFICIENT)	SEPERATOR VARIABLE	CODE LENGTH VARIABLE			
EXPONENT OF z		NUMBER OF UNIT BRANCHES TRAVERSED IN SERIES (LENGTH OF THE PATH)	NUMBER OF SYMBOLS GENERATED (LENGTH OF CODE WORD)	LENGTH OF DELAY	LENGTH OF DELAY	NUMBER OF STATE TRANSITIONS
GENERATING FUNCTION $G_{ik}(z)=\sum_{n=0}^{\infty} g_n z^n$	SOLUTION FOR x_k/Xi	g_n=NUMBER OF PATHS FROM NODE i TO NODE k WHOSE LENGTHS ARE EXACTLY n UNITS	g_n=NUMBER OF CODE WORDS OF EXACTLY n SYMBOLS POSSIBLE UNDER THE GIVEN SPECIFICATIONS	g_n=AMPLITUDE OF THE OUTPUT SIGNAL AT TIME $t=n$ DUE TO IMPULSE AT NODE i AT $t=0$.	g_n = OUTPUT OF ADDER k AT TIME $t=n$ DUE TO AN IMPULSE AT ADDER i AT $t=0$.	g_n = CONDITIONAL PROBABILITY OF THE SYSTEM OF BEING IN STATE k AT TIME $t = n$ GIVEN THAT THE SYSTEM WAS IN STATE i AT $t = 0$.

FIGURE 5

A GRAPH THEORETIC APPROACH TO LINEAR CODES

W. D. Frazer

Thomas J. Watson Research Center
International Business Machines Corporation
P. O. Box 218
Yorktown Heights, New York

In the years since Hamming's original paper, treatments of the theory of linear error correcting codes have, with a few notable exceptions, tended to adopt the viewpoint of group theory. It is the object of this note to discuss a somewhat different approach--that of linear graph theory*. Although the power of this approach is limited, it does admit some interesting insights and leads quite naturally to new classes of codes. Central to the exposition are the following well-known results and definitions**.

GROUND WORK AND BASIC THEOREMS

A binary linear code is a subspace of the vector space of n-tuples over the field mod 2; the n-tuples of the subspace are code words. The dimension, k, of the subspace is equal to the number of code word bits which carry information. The remaining (n-k) bits specify a set of parity checks, or linear dependence relations, among the k information bits of the code. The (Hamming) weight of a code word is defined to be the number of its nonzero positions, and the weight of a code to be the minimum weight of its constituent code words. An (n, k, d) code is, thus, a k-dimensional subspace of weight d of the space of n-tuples over the field mod 2.

---

*A colleague who has received a pre-publication copy has pointed out to the author that D. A. Huffman obtained independently some of the results herein, and will have reported on them at the ICMICT in Tokyo by the time this report appears.

**The notation used here will be that of Peterson (5) for coding theory and Seshu and Reed (7) for graph theory where no conflict arises. The reader is referred to these works for more details.

We shall be concerned in the sequel with loop-free, nonseparable, undirected linear graphs in which each pair of vertices is joined by at most one edge. The circuit matrix, $B = \left[b_{ij} \right]$, of such a graph has a column for each edge and a row for each circuit of the graph. An element, b_{ij}, of B is equal to one if edge e_j is part of circuit c_i and zero otherwise. The nullity, η, of such a graph of e edges and v vertices is $(e-v+1)$. The set of linear combinations (mod 2) of rows of the circuit matrix is a linear vector space with dimension equal to the nullity of the graph. This space, the vectors of which are circuits and unions of disjoint circuits, will hereafter be called the circuit space of the graph. The girth of a graph is the length of the shortest circuit.

As a natural consequence of these preliminaries, we assert: Theorem 1: The circuit space of a linear graph of e edges, v vertices, and girth g is an (e, η, g) code.

Now, any set of vectors which constitute a basis for a subspace can be taken as the rows of a matrix. Such a matrix is known as a generator matrix for a linear code which is the subspace. If we consider such a matrix, B_o, for a circuit space, and then transform the matrix to reduced echelon form: $B_o \rightarrow B_f = \left[I_\eta , B_{f_{12}} \right]$, where I_η is the unit matrix of order $\eta = (e-v+1)$ and $B_{f_{12}}$ is an $(e-v+1) \times (v-1)$ matrix, we have what is known as a fundamental circuit matrix for the graph. B_f is a generator matrix for the code which is the circuit space of the graph. The null space of a circuit space consists of vectors which are cut sets or disjoint unions of cut sets, and a matrix, $Q_f = \left[Q_{f_{11}}, I_{v-1} \right]$, whose rows are a basis for this null space is a fundamental cut set matrix. The edges corresponding to the columns of $B_{f_{12}}$ of B_f and to I_{v-1} of Q_f comprise a tree of the graph. (The columns of B_f and Q_f are assumed to be in the same order.) A matrix whose rows form a basis for the null space of a code is a parity check matrix. Thus:

Theorem 2: Any fundamental cut set matrix for a linear graph is a parity check matrix for the circuit space code of the graph.

CODES FOR COMPLETE GRAPHS

We now know how to generate a code from any linear graph, and can, therefore, investigate the codes corresponding to important kinds of graphs. The restrictions which we have imposed on the graphs under consideration require that any circuit have length at least three, and, therefore, that any circuit space code have distance at least three. No graph with these properties can have more edges than the complete graph, in which there is an edge joining every pair of vertices; therefore:

Theorem 3: The circuit space of the complete graph with v vertices is a $(\frac{v(v-1)}{2}, \frac{v(v-3)+2}{2}, 3)$ code. Furthermore, such a code has maximum transmission rate, $r = k/n$, among all circuit space codes for graphs of v vertices.

Proof: For the complete graph, $e = \frac{v(v-1)}{2}$

$$\eta = (e-v+1) = \frac{v(v-1)}{2} - v+1$$

$$= \frac{v(v-3)+2}{2}$$

By Theorem 1, the code is $(\frac{v(v-1)}{2}, \frac{v(v-3)+2}{2}, 3)$. Clearly, $r = \frac{e-(v-1)}{e}$ will be maximized for fixed v if e is maximized.

There are codes with the same minimum distance and greater transmission rate which are not the circuit spaces of linear graphs. The codes described in Hamming's original paper are such codes, and indeed the Hamming (7, 4, 3) code is the classic example, first cited by Whitney$^{(11)}$, of a vector space over the field mod 2 which is not realizable as the circuit space of any graph. Binary linear codes are in fact the circuit spaces of matroids over the field mod 2 and thus are not necessarily realizable as circuit spaces of linear graphs. The conditions under which such matroids are realizable as linear graphs have been defined by Tutte (9, 10). In order to pursue further the question of how circuit space codes for complete graphs compare with other linear codes, we digress briefly:

Since a linear code is a subspace of the vector space of n-tuples over the field, it is a subgroup of the additive group of n-tuples. One can, therefore, expand the entire group in cosets modulo the code. Further, since the choice of a leader is arbitrary within each coset, it is conventional to choose an element of minimum weight as coset leader. It can be shown that, for the binary symmetric channel, the average probability of correct decoding is maximized if a coset expansion of this form is used as a decoding table (5). Under these conditions, for a binary symmetric channel with error probability p and $q = (1-p)$, the probability of correct decoding of a received word can be written:

$$P\ (\text{correct decoding}) = \sum_{i=0}^{n} \alpha_i\, p^i\, q^{(n-i)} \qquad (1)$$

where α_i = number of coset leaders of weight i. Since (assuming $q > p$) the quantities $(p^i q^{n-i})$ decrease as i increases, the probability of correct decoding is increased whenever one α_i is increased and another later in the sequence is decreased. (Recall: $\sum \alpha_i = 2^{n-k}$, the number of cosets.) A code for the binary symmetric channel is called perfect if the coset expansion above has, for some m, all n-tuples of weight m or less and no others as coset leaders, and quasi-perfect if it has all patterns of weight m or less, some of weight $(m+1)$ and none of greater weight. A code is called optimum for the binary symmetric channel if its probability of error is as small as for any code with the same n and k. Clearly, perfect and quasi-perfect codes are optimum.

One might expect that the circuit spaces for complete graphs would be optimum codes, but this is again not the case, at least in general. For these codes, all n-tuples of weight one are coset leaders (thus necessary and sufficient for the code to have weight 3 or more) but not all n-tuples of weight 2 are coset leaders. This is easily seen: Consider a circuit of length 3 consisting of branches a, b, c. The n-tuple for this circuit will be in the code and the weight two n-tuples (a, b), (a, c), (b, c) will be in the cosets led by the weight one n-tuples c, b, a, respectively. Thus, the n-tuples corresponding to all pairs of branches with a common vertex will appear in cosets led by elements of weight one. Now in the complete graph, any pair of branches not having a common vertex are "opposite sides" of a circuit of length 4. Thus if such a pair of branches is chosen as a coset leader of weight two, the other two pairs of opposite sides will be in the same coset, since the circuit n-tuple is in the code. A detailed analysis of the above arguments will show that, for a code which is the circuit space of the complete graph with v vertices:

$$\text{no. wt. 2 n-tuples in wt. 1 cosets} = \frac{v\,!}{(v-3)\,!\;2}$$

$$\begin{array}{l}\text{no. wt. 2 n-tuples in wt. 2 cosets} \\ \text{(counting leaders)}\end{array} = \frac{v\,!}{(v-4)\,!\;6}$$

This accounts for all n-tuples of weight 2.

We can, therefore, compute the first three "error coefficients", α_i, of equation (1) for a code which is the circuit space of the complete graph with v vertices:

$$\alpha_0 = 1$$

$$\alpha_1 = n = \frac{v(v-1)}{2}$$

$$\alpha_2 = \frac{v\,!}{(v-4)\;4\,!}$$

The (trivial) code which is the circuit space of the complete graph with 3 vertices is perfect, and those for 4 and 5 vertices are quasi-perfect. (It may be verified by the reader that, for v = 4 or

5, $(a_0 + a_1 + a_2) = 2^{v-1}$). The code for 6 vertices (15, 10, 3) is not quasi-perfect, but is quite interesting because its transmission rate falls between that of the (15,9, 3) quasi-perfect code found by Fontaine and Peterson (1) and the (15, 11, 3) Bose-Chaudhuri code (5). It is not known whether this code is optimum, but those for 7 and 8 vertices are definitely not optimum. Fontaine and Peterson have found optimum codes which are better in both cases, as in Table 1.

	Complete Graph				Fontaine and Peterson		
Code	v	a_0	a_1	a_2	a_0	a_1	a_2
(21, 15, 3)	7	1	21	35	1	21	42
(28, 21, 3)	8	1	28	70	1	28	99

Table 1

The optimality of codes for larger values of v remains an open question.

CODES FOR BIPARTITE GRAPHS

A bipartite graph is one in which the vertices may be partitioned into two sets, Σ_1 and Σ_2, such that all branches connect vertices in one set to vertices in the other set. A complete bipartite graph is one in which each vertex in one set is connected to each vertex in the other. The number of branches in a complete bipartite graph is $|\Sigma_1| \cdot |\Sigma_2|$, where $|\Sigma_i|$ is the number of vertices in the set Σ_i. Since the girth of a bipartite graph is always at least 4, we have:

Theorem 4: The circuit space of a bipartite graph is a code of weight at least 4; the circuit space of a complete bipartite graph with vertex sets Σ_1 and Σ_2 is a:

$$(|\Sigma_1|\cdot|\Sigma_2|,\ |\Sigma_1|\cdot|\Sigma_2|-(|\Sigma_1|+|\Sigma_2|-1),\ 4) \text{ code.}$$

BOUNDS ON TRANSMISSION RATE

An interesting feature of binary linear codes which are the circuit spaces of linear graphs is their susceptibility to analysis by purely topological means. As an example of this phenomenon, consider the familiar Euler-Poincare characteristic, N. Let P be a polyhedron with e edges, v vertices, and f faces. Then $N_P = e-v+f$. Since such a polyhedron defines a surface, the Euler-Poincare characteristic of the surface defined by P is N_P (6). Now it can be shown that for any graph embeddable in a surface of characteristic N, $e \leq 3(v-N)$ (6). Using this result and the result of Theorem 1, some simple algebra yields:

Theorem 5: For any graph embeddable in a surface of characteristic N, the transmission rate, r, of the corresponding circuit space code is bounded as follows:

$$r \leq \frac{2}{3} - \left(\frac{N-1}{n}\right) . \qquad (2)$$

Note that, since N may be negative, this does not mean that the rate is always less than or equal to 2/3.

A regular graph is one in which the same number (called the degree or valence of the graph) of edges is incident at each vertex. If $L = \{l_{ij}\}$ is the set of paths between all pairs (i, j) of vertices of a graph, then $\delta = \max_{l_{ij}} (\min_{i,j} (l_{ij}))$ is the diameter of the graph. It can be shown (2) that in a graph which is regular and of degree t, the number of vertices is bounded above by:

$$v \leq 1 + t \sum_{i=1}^{\delta} (t-1)^{i-1} \qquad (3)$$

A little algebra leads us to the following bound on transmission rate:

$$v-1 \leq t\left(\sum_{i=1}^{\delta} (t-1)^{i-1}\right)$$

$$k = \eta = e-v+1 \geq e-t\left(\sum_{i=1}^{\delta} (t-1)^{i-1}\right)$$

$$\frac{k}{n} \geq 1-\frac{t}{n}\left(\sum_{i=1}^{\delta} (t-1)^{i-1}\right) \tag{4}$$

This bound is applicable to all codes which are the circuit spaces of regular graphs.

CODES DERIVED FROM CUBIC GRAPHS

A cubic graph is a regular graph of degree 3. Tutte (8) made an intensive investigation of the existence of cubic graphs of given girth having a minimum number of vertices, and exhibited the (unique) minimal graphs of girths 1 to 8, excluding 7. A minimal cubic graph of girth 7 was later found by McGee (4). The minimal cubic graph of girth 3 is the complete graph with 4 vertices; that of girth 4 is the Kuratowski hexagon (3); that of girth 5 is the Petersen graph (3). The minimal cubic graph of girth six is the map of seven hexagons which, when drawn on the torus, gives a map which cannot be colored with less than seven colors (6). The graph of girth eight is described in Tutte's paper. The codes corresponding to these graphs are, respectively, (6, 3, 3), (9, 4, 4), (15, 6, 5), (21, 8, 6) and (45, 16, 8). The code for McGee's cubic graph is (36, 13, 7).

MOORE GRAPHS

Regular graphs for which equality holds in (3) were first investigated by E. F. Moore. In such graphs, the girth is $(2\delta+1)$. Hoffman and Singleton (2) have investigated the existence of such graphs for the cases of girth 5 and 7. They found that Moore graphs of girth 5 exist only for $t = 2, 3, 7$ and (possibly) $t = 57$.

The unique Moore graph of girth 7 is a circuit (t=2). The cubic Moore graph of girth 5 is the Petersen graph mentioned previously. The code for the graph of degree 7 and girth 5 is (175, 126, 5). These codes are of interest mainly because they are "worst case" in transmission rate among codes which are the circuit spaces of regualr graphs of given degree and girth; this can easily be seen from (4).

CUT-SET SPACES AS CODES

Since the cut-set space of a graph is also a binary linear code (for which the circuit space is the null space), it is of interest to inquire whether cut set spaces might yield better codes. While no complete answer has been found, we can make some remarks about where to look.

Suppose one has a linear graph whose circuit space is an (e, η, d) code. One would then expect that if this code has a high rate the corresponding cut-set space code might not, in general, be a good code. In at least a restricted sense, one can demonstrate that this is so. Suppose we ask whether the corresponding cut-set code can have a higher rate while maintaining the same weight. Then we require:

$$\frac{\eta}{v-1} < 1$$

$$\frac{e-v+1}{v-1} < 1$$

$$e < 2v - 2$$

Now since the set of edges incident at any vertex is a cut-set, all vertices must have weight at least d. Thus:

$$dv \leq 2e$$

or

$$\frac{dv}{2} \leq e$$

Combining these results:

$$\frac{dv}{2} < 2v-2$$

$$d < 4-4/v < 4$$

Since the cases $d = 1, 2$ correct no errors and are, therefore, of limited interest, we see that only in the case of cubic graphs can a cut-set space code hope to complete on equal terms with the corresponding circuit space code. If we want to find a cut-set space code to compete with a circuit space code for $d > 3$, we must look for a different graph. The cut-set space of a complete graph is an $(e, v-1, v-1)$ code; thus it has much lower rate but greater weight (for $v > 4$) than the corresponding circuit space code.

REFERENCES

1. Fontaine, A. B., and Peterson, W. W.; "Group Code Equivalence and Optimum Codes", IRE PGIT-5 (1959) (Special Supplement), pp. 60-70.

2. Hoffman, A. J., and Singleton, R. R.; "On Moore Graphs with Diameters 2 and 3", IBM Journal of Research and Development, Vol. 4, No. 5 (November, 1960), pp. 497-504.

3. Konig, D., Theorie der Graphen; New York; Chelsea, 1950.

4. McGee, W. F.; "A Minimal Cubic Graph of Girth Seven"; Can. Math. Bull., Vol. 3, No. 2 (May, 1960), pp. 149-152.

5. Peterson, W. W.; Error Correcting Codes; Cambridge, Mass; MIT Press, 1961.

6. Ringel, G.; Farbungsprobleme auf Flachen und Graphen; Berlin: VEB Deutsche Verlag der Wissenschaften, 1959.

7. Seshu, S., and Reed, M. B.; Linear Graphs and Electrical Networks; Reading, Mass.: Addison-Wesley, 1961.

8. Tutte, W. T.; "A Family of Cubical Graphs"; Proc. Camb. Phil. Soc., Vol. 43 (1947), pp. 459-474.

9. Tutte, W. T.; "Matroids and Graphs"; Trans. Amer. Math. Soc., Vol. 90, (March, 1959), pp. 527-552.

10. Tutte, W. T.; "From Matrices to Graphs"; Can. Math. Journ., Vol. 16, No. 1 (1964), pp. 108-127.

11. Whitney, H.; "On the Abstract Properties of Linear Dependence", Amer. Journ. Math., Vol. 57 (1935), pp. 509-533.

ERROR CORRECTION AND CIRCUITS ON THE n-CUBE

R. T. Chien*+
C. V. Freiman*
D. T. Tang*

I. INTRODUCTION

The problem considered in this paper may be stated as follows: Given a unit or Boolean n-cube, and the graph formed by its edges, find the longest simple circuit in the graph such that the Hamming distance** d between any pair of vertices (codepoints) a and b on the circuit satisfy the relationship

$$d \geq \min (d_p, d_m) , \qquad (1)$$

where d_p is the shorter of the path distances between a and b on the circuit and d_m is the prescribed minimum distance, an integer.

The n-tuples corresponding to the vertices on the circuit form a code which will be called a circuit code.

The solution of this problem is useful in several important areas such as digital transmission of analog waveforms and design of asynchronous computer logic. A detailed discussion on applications of such codes has been given by Kautz.(3)

---

* IBM Watson Research Center, Yorktown Heights, N.Y.

+ Presently on leave at the University of Illinois.

** I.e., the number of positions in the two n-tuples at which their values differ.

When $d_m = 1$, the codes are generally known as Gray or unit-distance codes. This case has received attention from many authors. Some of the representative works are Tompkins,(4) Gilbert,(2) and Cohn.(1) Kautz has considered the case $d_m = 2$ and obtained upper bounds on the code length plus a table of codes up to $n = 7$.

In this paper, upper bounds on code size are derived for any d_m. Also presented are methods of construction which give forth classes of codes including some optimum codes.

II. AN UPPER BOUND

Consider a circuit code of minimum distance $d_m = 2t + 1$, dimension $n \geq d_m$ and length N*. We now proceed to show that

$$N \leq \frac{2^n}{\binom{n}{t} - 2\binom{n-1}{t-1}} = \frac{2^n}{\binom{n}{t}\left(1 - \frac{2t}{n}\right)} . \qquad (2)$$

This is equivalent to saying that a minimum of

$$\binom{n}{t} - 2\binom{n-1}{t-1} \qquad (3)$$

points must be uniquely assignable to each code point.

Consider a path of N points on the Boolean n-cube. We may enclose each code point with a Hamming sphere of radius t. As these spheres overlap, we seek to define a portion of each codeword's sphere such that:

i) this definition is independent of the codeword under consideration;

ii) portions corresponding to different codewords are disjoint; and

iii) the union of all portions is the same as the union of the original spheres.

---

* It is a simple matter to show that circuit codes of length $N = 2n$ exist independent of d_m and hence all subsequent arguments will only apply to cases where $N > 2n$.

One way of defining such sphere portions is to designate the codewords

$$a_0\ a_1\ a_2\ .\ .\ .\ a_{N-1}\ ,$$

preserving the order of codewords established by the circuit, and associated with each codeword a_i those points within its radius-t sphere which are not within the spheres of codewords $a_{i+1}\ a_{i+2} \ldots a_{i+2t}$ (all subscripts taken modulo N). Note that we have not included the spheres associated with codewords of the form a_{i+j} $(2t < j \leq N/2)$ as these t-radius spheres can never intersect the t-radius sphere about a_i without violating (1) for codewords a_i and a_{i+j}.

Without loss of generality, we assume that $i = 0$ and that a_0 is the code point with coordinates,

$$(0,\ 0,\ 0,\ \ldots\ ,\ 0),$$

the origin of the Boolean n-cube. This follows from the observation that distances do not change with translation. With suitable permutation of coordinates the sequence of codewords $a_0\ a_1 \ldots a_{2t}$ can be written as follows.

$$\begin{array}{ll}
a_0 = (0\ 0\ 0\ 0\ 0\ 0\ .\ .\ .\ .\ .\ .\ \ 0\ 0\ 0\ 0\ 0) & \\
a_1 = (0\ 0\ 0\ 0\ 0\ 0\ .\ .\ .\ .\ .\ .\ \ 0\ 0\ 0\ 0\ 1) & \\
a_2 = (0\ 0\ 0\ 0\ 0\ 0\ .\ .\ .\ .\ .\ .\ \ 0\ 0\ 0\ 1\ 1) & \qquad (4) \\
a_3 = (0\ 0\ 0\ 0\ 0\ 0\ .\ .\ .\ .\ .\ .\ \ 0\ 0\ 1\ 1\ 1) & \\
a_4 = (0\ 0\ 0\ 0\ 0\ 0\ .\ .\ .\ .\ .\ .\ \ 0\ 1\ 1\ 1\ 1) & \\
\qquad .\ .\ .\ .\ .\ .\ .\ .\ . & \\
a_{2t} = (0\ 0\ 0\ 0\ 0\ 0\ .\ .\ .\ \underbrace{1\ 1\ .\ .\ .\ 1\ 1\ 1\ 1}_{2t}) &
\end{array}$$

In other words, to fulfill the minimum distance requirement, it is necessary for the changing coordinates to be 2t different within this span of codepoints.

We shall now proceed to count the number of points that are 1) in the Hamming sphere of a_0, and 2) not in the Hamming sphere of a_i $(i = 1, 2, 3, \ldots, 2t)$. To obtain such a count we observe that each point which is of weight* t-1 or less is automatically in some sphere of a_i (i = 1, 2, ..., 2t). Furthermore, as the weights of a_i change in steps of one, if a codepoint b is of distance t to a_0 and of distance $\leq$ t to some $a_i (1 \leq i \leq 2t)$, there is always an $a_j (1 \leq j \leq 2t)$ such that the distance between b and a_j will be exactly t. Hence the number of points assigned to codeword a_0 is

$$\binom{n}{t} - B \tag{5}$$

where B is the number of elements in β, the set of points of weight t which are distance t from at least one $a_i (1 \leq i \leq 2t)$.

Consider any point $b \varepsilon \beta$, and let 2j be the smallest value of i for which b and a_i are of distance t —— it is clear that any point distance t from b must be of even weight as b is of weight t. From (4), we see that a_{2j} contains "1's" in its 2j rightmost coordinates. Exactly j of point b's "1's" must, therefore, be found in the 2j rightmost coordinates of b if a_{ij} and b are to be of distance t. We define $\beta_0 (\beta_1) \subset \beta$ to consist of those points in β with a "0" ("1") in their rightmost coordinate. It is clear that by complementing the 2j rightmost coordinates, a one-to-one mapping is established between β_0 and β_1. Taking B_1 to be the number of elements in β_1, we thus have: B equals $2B_1$.

Each point in β_1 must have a "1" in its rightmost coordinate and t-1 "1's" in the remaining n-1 coordinates. There are a total of $\binom{n-1}{t-1}$ points of this form. We shall now show that every point of this form is a point in β_1. Take any point of weight t with a "1" in its rightmost coordinate. Let e_i be the number of "1's" in the i rightmost coordinates less the number of "0's" in these coordinates. It is clear that e_1 = 1 and that

---

* The weight of n-tuble is the number of "1's" in its vector representation.

$e_{2t} = 0$. Therefore, there exists a j such that $e_{2j} = 0$ where $j \leq t$. This shows that any such point is in β_1 and consequently

$$B = 2B_1 = 2 \binom{n-1}{t-1} . \tag{6}$$

This leads immediately to

$$N \leq \frac{2^n}{\binom{n}{t} - 2\binom{n-1}{t-1}} = \frac{2^n}{\binom{n}{t}\left(1 - \frac{2t}{n}\right)} . \tag{7}$$

For the case $d_m = 2t$, we consider a "hard shell" of radius $t-1$ and a "soft shell" of radius t. The number of points distance t from a codepoint and not in the hard shell of any codepoint can be shown to be

$$S = \sum_{\alpha=1}^{t} \left[\binom{t+\alpha}{\alpha} - \binom{t+\alpha}{\alpha-1}\right] \sum_{\beta=\alpha}^{t} \binom{n-3t-2+\beta-\alpha}{\beta-\alpha} \left[\binom{2t+1-\beta}{t} - \binom{2t+1-\beta}{t+1}\right] \tag{8}$$

for $n > 3t$. Each of these points can be exactly distance t from (at most) $N/2$ codepoints and, hence,

$$N \leq \frac{2^n}{\frac{S}{N/2} + \binom{n}{t-1} - 2\binom{n-1}{t-2}} \tag{9}$$

or

$$N \leq \frac{2^n - 2S^{*}}{\binom{n}{t-1} - 2\binom{n-1}{t-2}} \tag{9}$$

III. CYCLIC CIRCUIT CODES

Because of the unit-distance property any circuit code can be efficiently represented by listing the changing coordinates rather than the sequence of code points. For instance, the code

---

* In the case of block codes of minimum distance $2t$, a point can be exactly distance t from at most $\lfloor n/t \rfloor$ codewords leading to the following bound or the number of codewords

$$\frac{2^n}{\binom{n}{t}/\lfloor n/t \rfloor + \sum_{a=0}^{t-1} \binom{n}{a}} .$$

$$\begin{array}{cccc}
x_1 & x_2 & x_3 & x_4 \\
\hline
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{array} \tag{10}$$

can be compactly represented by the sequence

$$4\ 3\ 2\ 1\ 4\ 3\ 2\ 1. \tag{11}$$

Such a sequence of N transitions shall be called the transition cycle* of the code sequence in that the last transition is taken to be adjacent to the first. From this point on we shall restrict our attention to this transition cycle as all distance properties of a circuit code may be directly determined from its corresponding transition cycle.

A circuit code is called cyclic if the changes of each coordinate occur in a periodic manner. Thus the circuit code of (11) is cyclic while the circuit code

$$4\ 3\ 2\ 1\ 4\ 2\ 3\ 1 \tag{12}$$

is not.

To facilitate the discussion we now introduce several definitions for cyclic circuit codes. Consider a cyclic circuit code of dimension n and length N. The order r_i of the ith coordinate is the number of occurrences of the ith coordinate in the transition cycle. The period of the ith coordinate in the transition cycle is taken to be $p_i = \frac{N}{r_i}$. We further define J to be the set of distinct values of r_i and ω_j to be the number of times j appears in the set of all r_i. The following elementary relations can then be easily verified.

---

* Tompkins(4) has used the term "flight sequence" for a similar type of representation.

$$2 \mid j \qquad j \in J$$

$$p_i r_i = N \qquad i = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} r_i = N$$

$$\sum_{j \in J} j \, \omega_j = N \tag{13}$$

$$\sum_{j \in J} \omega_j = n$$

Furthermore, the distance between two code points is easily obtained from the transition cycle by determining the number of coordinates which appear an odd number of times in either of the two transition cycles which carry one code point to the other. This observation coupled with the minimum distance requirement (1) immediately leads to

$$p_i > d_m , \quad i = 1, 2, \ldots, n, \tag{14}$$

which as a coordinate may not repeat until $d_p = d_m$. Also, since the circuit code corresponds to a simple circuit, each coordinate appears an even number of times in the transition cycle and the cycle is the shortest sequence of transitions for which this is true.

IV. A REALIZABLE CLASS OF CODES

In order to generate the class of codes we shall subsequently introduce, it is necessary first to establish the following theorem.

Theorem 1. Let

$$U = u_1 \, u_2, \ldots, u_{Nu}$$

$$V = v_1 \, v_2, \ldots, v_{Nv}$$

be the transition cycles of two cyclic circuit codes of minimum distance d_m, and dimensions n_u and n_v, respectively. If

$$m \equiv \mathrm{GCD}(N_u, N_v) > d_m, \tag{15}$$

then

$$u_1 \, v_1 \, u_2 \, v_2 \, u_3 \, v_3 \cdots$$

generates a cyclic circuit code with minimum distance d_m, dimension $n_u + n_v$, and of length $N_c = 2N_uN_v/m$.

Proof. Let us denote the transition cycle of the new code by

$$C = c_1 c_2 \, . \; . \; . \; c_i \; . \; . \; . \; c_{N_c}.$$

We note that the smallest value of i for which all $n_u + n_v$ coordinates appear an even number of times is given by

$$i = 2 \text{ LCM } (N_u, N_v) \equiv N_c ,$$

or

$$N_c = \frac{2 N_u N_v}{m} . \tag{16}$$

To prove that the constructed code satisfies (1), we will show that any sub-sequence

$$C_k = c_{i+1} \; c_{i+2} \; . \; . \; . \; c_{i+k} \quad (k \le N_c/2)$$

of C establishes a Hamming distance of at least d_m for $k > 2d_m$. (If $k \le 2d_m$, it is implied by the properties of U and V that the Hamming distance established is k.) Let $k_u(k_v)$ denote the number of times the elements of U(V) appear in C_k. Note that $k_u + k_v = k$ and $|k_u - k_v| \le 1$. We define reduced path distances R_u and R_v as follows:

$$k_u \equiv R_u \pmod{N_u}$$

$$k_v \equiv R_v \pmod{N_v}.$$

The Hamming distance d established by C_k is the Hamming distance d^u established by the elements from U plus the Hamming distance d^v established by the elements from V. Therefore,

$$d = d^u + d^v ,$$

and

$$d \ge \min (R_u, N_u - R_u, d_m) + \min (R_v, N_v - R_v, d_m).$$

Thus, $d \ge d_m$, unless

$$\min (R_u, N_u - R_u) + \min (R_v, N_v - R_v) < d_m. \tag{17}$$

To show that this is impossible, we first investigate the case in which (17) takes the form

$$R_u + R_v < d_m. \tag{18}$$

Let j_u be the largest integer $\le i + k$ such that: C_{i+j_u}

is an element of U; and the number of occurences of elements of U in the sequence $c_{i+1}\ c_{i+2}\ \cdot\ \cdot\ \cdot\ c_{i+j_u}$ is a multiple of N_u. Note that this implies that $R_u = \left[\frac{k-j_u}{2}\right]$*. We define j_v in an analogous manner. Then it follows that $|j_u-j_v| \equiv 1 \pmod{2m}$ since m is the GCD(N_u,N_v). If $|j_u-j_v|$ is 2m-1 or greater, then $R_u + R_v \geq m-1 \geq d_m$--- contradicting (17). Thus $|j_u-j_v|$ must equal 1 if (17) is to be satisfied. This implies that j_u and j_v are i+1 and i+2 in some order. Thus, $k = R_u + R_v < d_m$ if (17) is to be satisfied but this corresponds to a case where Hamming distance equals path distance as was observed earlier.

An exactly parallel argument holds when (17) takes the form

$$N_u - R_u + N_v - R_v < d_m.$$

As u and v are interchangeable, the last case we need consider is that in which (17) takes the form

$$R_u + N_v - R_v < d_m. \qquad (19)$$

In this case, j_u is defined as above, but we define j_v' to be $j_v' = j_v + 2N_v$. We note that the smallest value of $j_v' - j_u$ is 2m-1 and we take $j_v' - k = x$, $h - j_u \geq 2m-1 - x$. From this,

$$d_v \geq \min\left(\frac{x}{2}, d_m\right),$$

$$d_u \geq \min\left(\frac{2m-1-x}{2}, d_m\right).$$

By examining all combinations it follows that

$$d_u + d_v \geq \min\left(d_m, \frac{2m-1-x}{2} + \frac{x}{2}\right)$$

$$= \min\left(d_m, \frac{2m-1}{2}\right)$$

$$\geq d_m.$$

This shows no sequence C_k exists for which (1) is not satisfied.

This theorem is easily used to generate new codes from codes of lower dimensions. An example of a class of codes constructable in this way is given by the following. For a specified minimum distance d_m, we first observe that codes with dimension n_0 and $N_0 = 2n_0$ codewords, having the following transition cycle, exist.

---

* We use [ ] to denote "integer part of".

$$1\ 2\ .\ .\ .\ n_0\ \ 1\ 2\ .\ .\ .\ n_0.$$

Applying Theorem 1 to two such codes of dimension n_0 will yield a code of dimension $n_1 = 2n_0$ and length $N_1 = 2N_0$ provided (15) is satisfied by having

$$\mathrm{GCD}(2n_0,\ 2n_0) = 2n_0 > d_m. \tag{20}$$

Theorem 1 is now applied to the resultant code and the original code yielding a new code with $n_2 = 3n_0$ and $N_2 = 8n_0$. Note that no additional constraint is placed on n_0 as (15) simply requires

$$\mathrm{GCD}(2n_0,\ 4n_0) = 2n_0 > d_m.$$

By repeating the above procedure, it is clear that circuit codes with the following parameters can be constructed:

$$\begin{aligned} &\text{dimension } n_k = (k+1)\ n_0, \\ &\text{length } \ N_k = 2^{k+1}\ n_0, \end{aligned} \tag{21}$$

for $k = 0, 1, 2, \ldots$, provided $2n_0 > d_m$.

V. ADDITIONAL PROPERTIES

The following theorems significantly reduce the effort necessary to find the optimum transition cycle for specified dimension and minimum distance.

Theorem 2. For any cyclic circuit code, the greatest common divisor of the r_i's is two.

Proof. Let $g = \mathrm{GCD}(r_i)$. From (13), $2|g$. Since $N = r_i p_i$, $\frac{2N}{g} = \frac{2r_i}{g}\ p_i$, and thus, within a span of $\frac{2N}{g} \leq N$ transitions, the ith coordinate has $\frac{2r_i}{g}$ transitions. As $g|r_i$, each coordinate undergoes an even number of transitions in this $\frac{2N}{g}$ span, thus forming a closed circuit of codewords. This implies $N \leq \frac{2N}{g}$ and, therefore, we must have $N = \frac{2N}{g}$ or $g = 2$.

Theorem 3. For all $j \in J$, the following condition holds.

$$\left(\frac{N}{j\omega_j} - 1\right)(2\omega_j - 1) \geq d_m. \tag{22}$$

Proof. Choose $j \in J$ and let $p = N/j$. The average spacing between adjacent occurrences of any coordinates with this cycle period is simply $\frac{p}{\omega_j}$. Therefore, there always exists a span of $\frac{p}{\omega_j} - 1$ transitions in which no coordinate with period p appears. A span of length $2p - (\frac{p}{\omega_j} - 1)$ containing exactly two appearances of each of these coordinates is thus seen to exist always. The largest distance which can be found in such a span is $2p - (\frac{p}{\omega_j} - 1) - 2\omega_j$, and hence (22).

Theorem 4. For a cyclic circuit code with minimum distance d_m.

$$\sum_{\substack{j \in J \\ 4 \times j}} \omega_j - d_m. \tag{23}$$

Proof. The left side of (23) is the distance established by the transition sequence at half the cicle. Hence the theorem.

A somewhat similar theorem is the following.

Theorem 5. For each $j \in J$ and any $k = 1, 2, \ldots$

$$2k\left(\frac{N}{j} - \omega_j\right) - 2\sum_i t_i \omega_i \geq d_m, \tag{24}$$

where $i \in J$, such that an integer t_i exists satisfying

$$(2t_i - 1)\frac{N}{j} < 2k\frac{N}{i} \leq 2t_i\frac{N}{j}.$$

Proof. The left side of (24) is an upper bound of the distance established by the transition sequence at length $= 2k\frac{N}{j}$. Hence the theorem.

Theorem 6. Let $i, j \in J$ $(i = j)$, $h_{ij} = \mathrm{GCD}\left(\frac{N}{i}, \frac{N}{j}\right)$ then

$$\left[\frac{i\omega_i h_{ij}}{N}\right]^{+ *} + \left[\frac{j\omega_j h_{ij}}{N}\right]^{+} \leq h_{ij} . \qquad (25)$$

Proof. Let us represent the transition cycle by $t_1 t_2 \ldots t_N$. As $i|N$ and $j|N$ it is clear that $h_{ij}|N$. We shall thus define t_e and t_f to be in phase if $e-f \equiv 0 \pmod{h_{ij}}$. It can be shown that a coordinate whose cycle period is $\frac{N}{i}$ and a coordinate whose cycle period is $\frac{N}{j}$ can never occupy positions of the same phase. Were this the case, the relative primeness of $\frac{N}{i\, h_{ij}}$ and $\frac{N}{j\, N_{ij}}$ would require that both coordinates occupy a single transition position. The total number of occurrences of coordinates whose cycle period is $\frac{N}{i}$ is simply $i\omega_i$. The smallest number of different phases occupied by such coordinates is

$$\left[\frac{i\omega_i}{N/h_{ij}}\right]^{+} .$$

Since the total number of phases is h_{ij}, we have (25).

VI. OPTIMUM CYCLIC CIRCUIT CODES

The following table lists the parameters of some existing optimum cyclic circuit codes

TABLE I.

Circuit Code Bounds and Optimum Cyclic Circuit Codes

n	$d_m=3$		$d_m=4$		$d_m=5$	
	Eq(2)	N_{opt}	Eq(9)	N_{opt}	Eq(2)	N_{opt}
2	-	4	-	4	-	4
3	-	6	-	6	-	6
4	8	8	-	8	-	8
5	10	10	10	10	-	10
6	16	16	14	12	12	12
7	24	24	22	14	14	14
8	42	32	38	20	18	16
9	72	48**	66	24**	24	24

---

* $[\ ]^{+}$ denotes "least integer not less than."

** The largest N obtained, not necessarily optimum.

To illustrate how much optimum code lengths can be determined, consider the case with n = 8 and $d_m = 3$. From (2) the upper bound is found to be 42. We investigate the solutions to the following equations.

$$\begin{cases} \sum\limits_{j \in J} \omega_j = n, \\ \sum\limits_{j \in J} j\, \omega_j = N. \end{cases} \tag{26}$$

For N = 42 = 2 x 3 x 7, the acceptable elements of J are 2 and 6 from (13) and (14). But there exists no solution to (26) with this J.

For N = 40 = 2 x 2 x 2 x 5, we have J = {2, 4, 8, 10}. Solutions to (26) are listed in the following table.

TABLE II.

Values of $\bar{\omega}_j$

j =	2	4	8	10	ruled out by
Case 1	5	0	0	3	Th. 3 on j = 10.
2	4	1	1	2	Th. 6 on j = 8, 10.
3	4	0	4	0	Th. 3 on j = 8.
4	3	2	2	1	Th. 6 on j = 8, 10.
5	2	4	0	2	
6	2	3	3	0	Th. 4.
7	1	5	1	1	Th. 6 on j = 8, 10.
8	0	6	2	0	Th. 2.

The only possibility is case 5 which yields a transition cycle

$$||:\ 1\ 3\ 2\ 7\ \overbrace{1\ 4\ 2\ 5\ 1\ 6\ 2\ 3\ 1\ 8\ 2\ 4\ 1\ 5\ 2\ 6}\ :||\ *$$

which fails the distance requirement over the span indicated.

For N = 38 = 2 x 19, we have J = {2} for which no solution to (26) exists.

For N = 36 = 2 x 2 x 3 x 3, we have J = {2, 4, 6}. Solutions to (26) are listed in Table III.

---

* We borrow the symbol ||: . . . :|| to mean "repeat."

TABLE III.

Values of $_j$

j =	2	4	6	ruled out by
Case 1	3	0	5	Th. 3 on j = 6.
2	2	2	4	Th. 5 on j = 6.
3	1	4	3	Th. 6 on j = 4, 6.
4	0	6	2	Th. 4.

For $N = 34 = 2 \times 17$, we have $J = \{2\}$, and again no solution to (26) exists.

An optimum code with $N = 32$ is obtained by applying Theorem 1 to the following codes of $d_m = 3$:

code 1 ($n_1 = 2$, $N_1 = 4$) : $\|$: 1 2 :$\|$

code 2 ($n_2 = 6$, $N_2 = 16$): $\|$: 3 5 4 6 3 7 4 8 :$\|$.

This yields

code 3 ($n_3 = 8$, $N_3 = 32$): $\|$: 1 3 2 5 1 4 2 6 1 3 2 7 1 4 2 8 :$\|$

REFERENCES

1. Cohn, M., "Affine M-ary Gray Code," Information and Control, Vol. 6. (1963) pp. 70-78.

2. Gilbert, E. N., "Gray Codes and Paths on the n-Cube," Bell System Tech. J., Vol. 37 (1958) pp. 179-180.

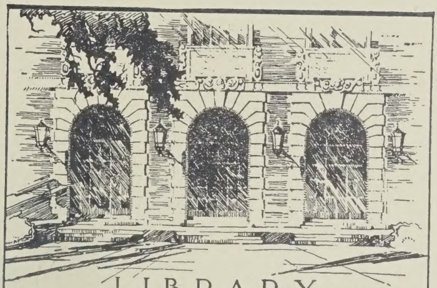
3. Kautz, W. H., "Unit-Distance Error-Checking Codes," I. R. E. Trans., Vol. EC-7 (1958) pp. 179-180.

4. Tompkins, H. E., "Unit Distance Binary Codes," Report No. 58-09, Moore School of Electrical Engineering, U. of Pa., Philadelphia (1957).

INDEX OF AUTHORS

Ahmed, N. 220
Anderson, D.R. 366
Arozullah, M. 394
Auth, L.V.,Jr. 288
Bedrosian, S.D. 70,81
Beene, R.H. 802
Bickart, T.A. 243,273
Biorci, G. 335
Boesch, F.T. 559
Brockett, R.W. 413
Brown, D.P. 431
Calahan, D.A. 175
Chen, W.H. 250
Chen, Y.C. 530
Cheng, K.Y. 135
Chien, R.T. 899
Civalleri, P.P. 293
Desoer, C.A. 1
Forys, L.J. 413
Frank, H. 485
Frazer, W.D. 888
Freiman, C.B. 899
Frisch, I.T. 467
Fu, K.S. 618
Glinski, G.A. 394
Haase, K.H. 745
Happ, W.W. 508
Kalman, R.E. 32,43
Karni, S. 220
Ku, W.H. 765
Lago, G.V. 573
Lamb, R.D. 508
Larson, R.E. 643
Lasso, M.A.Murray 820
Layton, D.M. 467
Lee, H.B. 447
Lee, H.T. 699
Leon, B.J. 366
Liou, M. -L 114
Liu, R. -w 228,864
Marshall, T.G.,Jr. 194
Massey, J.L. 864
McMurtry, G.J. 618
Meditch, J.S. 666
Meo, A.R. 841
Merkl, E.D. 802
Murphy, G.J. 46
Murray Lasso, M.A. 820
Myers, B.R. 544
Peschon, J. 596
Pottle, C. 201
Pullen, K.A. 734
Ramamoorthy, C.U. 875
Reis, G.C. 100,233
Robbins, E.R. 508
Rogers, R.H. 431
Rosenthal, C.W. 337
Sablatash, M. 149
Sage, A.P. 682
Seacat, R.H. 802
Shenoi, B.A. 782
Simon, H. 337

Streets, R.B. 682
Suffel, C.L. 559
Tang, D.T. 83,899
Thorp, J.S. 201
Weinberg, L. 42,81
Weiss, L. 15,377
Wing, O. 530
Wu, S.H. 46
Zobrist, G.W. 573



LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
2 1964
3 0112 006279597